Bemba Speech Translation: Exploring a Low-Resource African Language

Muhammad Hazim Al Farouq
Kreasof AI
Research Labs
Jakarta, Indonesia

Aman Kassahun Wassie
African Institute for
Mathematical Sciences (AIMS)
Addis Ababa, Ethiopia

Yasmin Moslem[☆]
ADAPT Centre
Trinity College Dublin
Dublin, Ireland

Abstract

This paper describes our system submission to the International Conference on Spoken Language Translation (IWSLT 2025), low-resource languages track, namely for Bemba-to-English speech translation. We built cascaded speech translation systems based on Whisper and NLLB-200, and employed data augmentation techniques, such as back-translation. We investigate the effect of using synthetic data and discuss our experimental setup.

1 Introduction

Low-resource languages face critical limitations due to the scarcity and scattered nature of the available data (Haddow et al., 2022). Speech translation for low-resource languages involves similar challenges (Ahmad et al., 2024; Moslem, 2024; Lovenia et al., 2024; Abdulmumin et al., 2025), Similarly, speech applications for African languages are very limited due to the lack of linguistic resources. For example, Bemba is an under-resourced language spoken by over 30% of the population in Zambia (Sikasote and Anastasopoulos, 2022). Hence, the IWSLT shared task on speech translation for low-resource languages aims to benchmark and promote speech translation technology for a diverse range of dialects and low-resource languages.

We participated in the Bemba-to-English language pair through building cascaded speech translation systems. In other words, we employed Whisper (Radford et al., 2022) for automatic speech recognition (ASR), and NLLB-200 (Costa-jussà et al., 2022) for text-to-text machine translation (MT). For ASR, we fine-tuned Whisper models using two datasets, BembaSpeech and BIG-C. For MT, we fine-tuned the NLLB-200 models using the bilingual segments of the BIG-C dataset, and the "dev" split of the FLORES-200 dataset. In addition, we augmented the Bemba-to-English training data with back-translation of a portion of the Tatoeba dataset from English into Bemba. The back-translated data was filtered based on cross-entropy scores. As Table 4 shows, the systems we submitted to the shared tasks are as follows:

- Primary: It uses Whisper-Medium for ASR and NLLB-200 3.3B for MT.
- Contrastive 1: It uses Whisper-Small for ASR and NLLB-200 3.3B for MT.
- Contrastive 2: It uses Whisper-Small for ASR and NLLB-200 600M for MT.

2 Data

The data we used to train our Bemba-to-English speech translation models can be categorized into: (1) authentic data, and (2) synthetic data. The following sections provide more details (cf. Table 1).

Dataset	Language	Train	Dev	Test	Audio
Big-C	Bem-Eng	82,371	2,782	2,763	✓
BembaSpeech	Bem	12,421	1,700	1,359	✓
FLORES-200	Bem-Eng	997	0	1,012	✗
Tatoeba	Eng	20,121	0	0	✗

Table 1: Data Statistics: The "Language" column specifies which languages are originally available in each dataset. "Train", "Dev", and "Test" represent the dataset sizes. The "Audio" column indicates whether each dataset includes audio signals.

2.1 Authentic Data

We filtered the authentic data by removing any overlaps between the training data and test data based on the text transcript. For building our models, we used the following data sources.

- **Big-C** is a parallel corpus of speech and transcriptions of image-grounded dialogues between Bemba speakers and their corresponding English translations. It contains 92,117

[☆]Correspondence: yasmin[at]machinetranslation.io

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 354–359
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

Training Dataset(s)	FLORES-200			BIG-C		
	BLEU	chrF++	COMET	BLEU	chrF++	COMET
Big-C	18.13	42.11	53.25	27.83	51.08	53.28
Big-C + Tatoeba	21.67	45.25	55.64	27.82	50.98	**53.39**
Big-C + FLORES-200	25.21	47.31	57.23	27.96	51.03	53.29
Big-C + FLORES-200 + Tatoeba	**25.70**	**47.75**	**58.29**	**28.60**	**51.38**	53.08

Table 2: MT Evaluation: In general, the models trained with both authentic data (Big-C & FLORES-200) and back-translated data (Tatoeba) outperform the models trained with the authentic data only. All the models in this table uses NLLB-200 600M.

spoken utterances of both complete and incomplete dialogues, amounting to 187 hours of speech data grounded on 16,229 unique images. The dataset aims to enable the development of speech recognition, speech, and text translation systems for Bemba, as well as facilitate research in language grounding and multimodal model development (Sikasote and Anastasopoulos, 2022).[1] Since this dataset includes audio and transcription in Bemba as well as translation into English, we could use it to build both modules of our cascaded systems, i.e. ASR and MT. Table 7 shows examples of sentence pairs from the Big-C datasets.

- **BembaSpeech** is an ASR corpus for the Bemba language of Zambia. It contains read speech from diverse publicly available Bemba sources; literature books, radio/TV shows transcripts, YouTube video transcripts as well as various open online sources. Its purpose is to enable the training and testing of automatic speech recognition (ASR) systems in Bemba language. The corpus has 14,438 utterances, culminating into 24.5 hours of speech data (Sikasote et al., 2023).[2] We used the BembaSpeech dataset in addition to the Big-C dataset to build our ASR models.

- **FLORES-200** (Goyal et al., 2022) is a bilingual text-only dataset for machine translation. We used the Bemba-to-English "dev" split for training, and the "devtest" split for testing.

- **Tatoeba** (Tiedemann, 2020) is a monolingual dataset in English. We used a portion of it for back-translation (cf. Section 2.2).

[1] https://github.com/csikasote/bigc

[2] https://github.com/csikasote/BembaSpeech

2.2 Synthetic Data

We augmented our authentic data (cf. Section 2.1) with synthetic data created with back-translation. To this end, we fine-tuned the NLLB-200 600M model in the other direction, i.e. for the English-to-Bemba language pair. Thereafter, we translated the English sentences from Tatoeba into Bemba using the fine-tuned English-to-Bemba NLLB-200 model. For translation, we used CTranslate2 (Klein et al., 2020), generating the prediction cross-entropy scores for each sentence, and calculating the exponential of the scores for better readability. We filtered data based on the cross-entropy scores, removing low-quality segments. We removed segments with scores less than 0.77 based on manual exploration of samples of the generated back-translations. While the unfiltered back-translated data consists of 85,000 segments, the filtered back-translated data consists of 20,000 segments. Finally, we prepended the source side (Bemba) with the <*bt*> tag to indicate that the data is synthetic. Moreover, we experimented with removing the <*bt*> tag and found that this achieves slightly better results when testing with the FLORES-200's "devtest" split, as the data was already filtered (cf. Table 3).

3 Experiments and Results

As illustrated by Figure 1, our cascaded systems involve two components, an ASR model based on Whisper to generate transcriptions and an MT model based on NLLB-200 to generate text translation. We experimented with different versions of these models, namely Whisper Small and Medium, and NLLB-200 with 600M and 3.3B parameters. Our code for data preparation, training, and evaluation is publicly available.[3]

[3] https://github.com/cobrayyxx/Bemba-IWSLT2025

Datasets	BT Size	Filtered	<bt> tag	FLORES-200 BLEU	FLORES-200 chrF++	FLORES-200 COMET	BIG-C BLEU	BIG-C chrF++	BIG-C COMET
	85,155	✗	✓	20.96	45.06	**55.92**	**28.17**	**51.26**	53.45
BIG-C + Tatoeba	20,121	✓	✓	19.82	44.09	54.79	28.04	51.20	**53.51**
	20,121	✓	✗	**21.67**	**45.25**	55.64	27.82	50.98	53.39

Table 3: Performance of MT models that are based on NLLB-200 600M and trained using both authentic data and augmented back-translated data. There are two pre-processing aspects applied to the augmented data, filtering the data based on cross-entropy scores, and prepending the source sentence with the <*bt*> tag. Evaluating the models with the devtest split of the FLORES-200 dataset, the highest evaluation scores, in terms BLEU and chrF++, are achieved when the back-translated data is filtered and the <*bt*> tag is removed. Meanwhile, the AfriCOMET score (COMET) of this model is comparable to the model where the back-translated data is not filtered and the source is prepended with the <*bt*> tag. Evaluating the models with the hold-out test split of Big-C reveals a different outcome where using the <*bt*> tag results in relatively higher scores, although the scores of the three experiments are relatively comparable. It is worth noting that the filtered back-translated data consists of only 20k segments, while the unfiltered back-translated data consists of 85k segments.

Training: We trained our models for 3 epochs, saving the best checkpoint based on the chrF++ score during training on the validation dataset. Our training arguments were chosen based on both manual exploration and automatic hyperparameter optimization using the Optuna framework (Akiba et al., 2019). The most important arguments are a learning rate of 1e-4 and a warm-up ratio of 0.03.

Inference: For inference, we used Faster-Whisper [4] with the default VAD[5] arguments, and 5 for the "beam size". The model was quantized with the float16 precision for more efficient inference.

Evaluation: To evaluate our systems, we calculated BLEU (Papineni et al., 2002), and chrF++ (Popović, 2017), as implemented in the sacreBLEU library[6] (Post, 2018). For semantic evaluation, we used AfriCOMET (Wang et al., 2024). We conducted ASR evaluation (cf. Table 5) and MT evaluations (cf. Table 2 and Table 3). Finally, we evaluated the whole cascaded systems (cf. Table 4).

3.1 Data Augmentation

As explained in Section 2.2, we created synthetic data using back-translation to augment our training data (Sennrich et al., 2016; Edunov et al., 2018; Poncelas et al., 2019; Haque et al., 2020). Then, we filtered this back-translated data based on generation cross-entropy scores. In our experiments, data augmentation improved the translation quality.

[4] https://github.com/SYSTRAN/faster-whisper

[5] Voice Audio Detection (VAD) removes low-amplitude samples from an audio signal, which might represent silence or noise.

[6] https://github.com/mjpost/sacrebleu

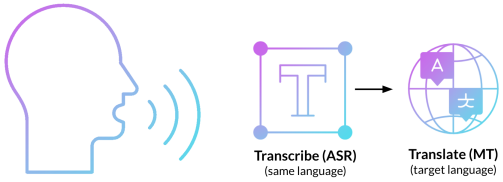

Figure 1: Cascaded speech translation systems use two models, an ASR model to generate audio transcriptions in the same language, and then an MT model to translate the generated transcriptions into the target language.

As shown in Table 2, when fine-tuning NLLB-200 600M, the models trained with back-translated data outperformed the models trained with only the authentic data.

We tried prepending the back-translated source with the <*bt*> tag, but found removing it achieves better results (cf. Table 3). This might be because we filtered the back-translated data, so its quality is good enough that it does not require distinguishing from the authentic data with the <*bt*> tag.

3.2 Whisper and NLLB-200 Models

We experimented with both Whisper Small and Whisper Medium to train ASR models. Similarly, we experimented with both NLLB-200 600M and 3.3B to train MT models. For our datasets, the results are comparable (cf. Table 4).

3.3 End-to-End vs. Cascaded System

Unlike a cascaded system, an end-to-end speech translation system requires only one model to perform audio-to-text translation (Agarwal et al., 2023;

System	ASR	MT	Type	BLEU	chrF++	COMET
Primary	Whisper-Medium	NLLB 200 3.3B	Baseline	0.72	14.28	16.23
			Finetuned	**27.45**	**49.64**	**51.74**
Contrastive 1	Whisper-Small	NLLB 200 3.3B	Baseline	0.51	13.41	11.9
			Finetuned	**27.39**	**49.65**	**52.01**
Contrastive 2	Whisper-Small	NLLB 200 600M	Baseline	0.41	13.21	10.69
			Finetuned	**27.30**	**50.17**	**51.91**

Table 4: Performance of the baseline and finetuned cascaded systems based on BLEU, chrF++, and AfriCOMET (COMET) scores. The approaches we followed, including fine-tuning and data augmentation, have considerably improved the quality of Bemba-to-English speech translation. The models were evaluated using the test split of the Big-C dataset.

Model	Type	WER
Whisper-Small	Baseline	157.5
	Finetuned	35.64
Whisper-Medium	Baseline	150.92
	Finetuned	**36.19**

Table 5: ASR Evaluation: The models were trained with Big-C and BembaSpeech. The performance of the finetuned models outperform the baseline models, indicated by the lower Word Error Rate (WER) scores of the finetuned models compared to the baseline models. The models were evaluated using the test split of the Big-C dataset.

Ahmad et al., 2024; Moslem et al., 2025). We finetuned Whisper directly on the Bemba-to-English Big-C dataset. Table 6 compares the results of the two systems. Where there is a slight increase in the scores of BLEU And chrF++ of the end-to-end model, the cascaded system outperforms the end-to-end system in terms of the COMET score, while the BLEU score of the end-to-end model is slightly higher.

Model	Type	BLEU	chrF++	COMET
End-to-End	Baseline	0.09	11.85	6.9
	Finetuned	**28.08**	**49.68**	48.36
Cascaded	Baseline	0.51	13.41	11.9
	Finetuned	27.39	49.65	**52.01**

Table 6: Comparison of the end-to-end speech translation using Whisper-Small, and the cascaded system that uses Whisper-Small for transcription and then NLLB-200 3.3B for translation. The evaluation uses the test split of the Big-C dataset.

🔊	nafwala na amakalashi ku menso
CC	he is wearing glasses as well
A文	He is wearing glasses.
🔊	Imbwa iyafonka pamoona, ilebutuka palunkoto lwamucibansa
CC	A dog with a wide nose is running on the lawns of the football ground
A文	A dog with a pointed nose is running on the lawn
🔊	Akamwanakashi nakemya ukuulu mumuulu ukulwisha ukutoba aka lipulanga.
CC	She has her leg in the air attempting to break a board.
A文	A child has lifted one leg in an attempt to hit a wood.
🔊	Kunuma yabo kuli notu ma motoka tulya ba bonfya mu ncende iya talala nge iyi baliko.
CC	There are also small vehicles that they use in cold places behind them.
A文	Behind them are vehicles that they use in cold places like this one.
🔊	abaume Bali pa mutenge yanganda umo afwele ishati lya mitomito ilyamaboko ayatali elyo me tolishi lya makumbimakumbi
CC	Of the men on the roof of the house, ine is wearing a long sleeved grey shirt and a blue trousers.
A文	Men are on the roof of the house, one is wearing a grey long sleeved shirt and a blue trousers.
🔊	Ifi bafwele kunsapato fyakutelelela nga baya mukwangala umu mwine muli ice.
CC	These on their shoes are for sliding when the to play on the ice.
A文	These shoes they are wearing are for sliding when they are going to play in ice.
🔊	Namayo ale enda mumusebo nabika nomwana pamabeya.
CC	A woman is walking in the road with a child on her showders.
A文	A woman is walking in the road with a child on her shoulder.
🔊	Namayo naikata ifyakulya pa mbale mukati ke tuuka.
CC	A woman is holding food on a plate inside a shop.
A文	A woman is holding food on a plate inside a shop.
🔊	Nangu limbi kuli bamo abamufulwishe.
CC	Or maybe someone has made him upset.
A文	Or maybe someone has upset her.
🔊	Abantu bane bali umuli ifimabwe ifikulu nga nshi kabili nafwala ne fimpopo ku mitwe yabo
CC	four people are inside an area with large rocks and they are wearing helmets
A文	Four people are in a place full of rocks and they are wearing helmets.
🔊	Akamwana kambi balekafuula amasapato kuli kafundisha wakako.
CC	Another child's shoes being taken off by the instructor.
A文	One of the pupils is being removed the shoes by the teacher.
🔊	Afwile alefwaya afike pampela ya lumpili. Pantu icishimbi ekete.Eco babomfya abatemwa ukuniine mpili
CC	Maybe he wants to reach the top of the mountain. The rode metal he is carrying, it is mostly used when one is climbing the mountains.
A文	Obviously he wants to reach the top of the mountain because this metal he is holding is used by mountain climbers.

Table 7: Examples of sentences in Bemba, their English translations from the Big-C dataset, and generated translations using Whisper-Medium and NLLB-200 3.3B.

Acknowledgements

We would like to thank Kreasof AI for supporting this work through providing the first author with computational resources.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Ashwin, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Marco Gaido, Dávid Javorský, Marek Kasztelnik, Tsz Kin Lam, Danni Liu, Evgeny Matusov, Chandresh Kumar Maurya, John P. McCrae, Salima Mdhaffar, Yasmin Moslem, Kenton Murray, Satoshi Nakamura, Matteo Negri, and 20 others. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austia (in-person and online). Association for Computational Linguistics.

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, and 33 others. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kim Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John McCrae, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Atul Kr Ojha, John Ortega, Sara Papi, Peter Polák, Adam Pospíšil, and 15 others. 2024. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Stroudsburg, PA, USA. Association for Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

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, and 9 others. 2022. No Language Left Behind: Scaling human-centered machine translation. *arXiv [cs.CL]*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Trans. Assoc. Comput. Linguist.*, 10:522–538.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Machine Translation. *Computational Linguistics*, 06:1–67.

Rejwanul Haque, Yasmin Moslem, and Andy Way. 2020. Terminology-Aware Sentence Mining for NMT Domain Adaptation: ADAPT's Submission to the Adap-MT 2020 English-to-Hindi AI Translation Shared Task. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): Adap-MT 2020 Shared Task*, pages 17–23, Patna, India. NLP Association of India (NLPAI).

Guillaume Klein, Dakun Zhang, Clément Chouteau, Josep Crego, and Jean Senellart. 2020. Efficient and high-quality neural machine translation with OpenNMT. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 211–217, Stroudsburg, PA, USA. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, and 32 others. 2024. SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages. *arXiv [cs.CL]*.

Yasmin Moslem. 2024. Leveraging Synthetic Audio Data for End-to-End Low-Resource Speech Translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 265–273.

Yasmin Moslem, Juan Julián Cea Morán, Mariano Gonzalez-Gomez, Muhammad Hazim Al Farouq, Farah Abdou, and Satarupa Deb. 2025. SpeechT: Findings of the first mentorship in speech translation. In *Proceedings of Machine Translation Summit XX, Implementations and Case Studies Track*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. Adaptation of Machine Translation Models with Back-Translated Data Using Transductive Data Selection Methods. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing CICLing 2019: Computational Linguistics and Intelligent Text Processing*, pages 567–579, La Rochelle, France. Springer Nature Switzerland.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv [eess.AS]*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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

Claytone Sikasote, Eunice Mukonde, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. BIG-C: a multimodal multi-purpose dataset for Bemba. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2062–2078, Toronto, Canada. Association for Computational Linguistics.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jiayi Wang, David Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Mohamed, and 29 others. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Stroudsburg, PA, USA. Association for Computational Linguistics.