Data statements for BembaSpeech Dataset

Data set name: Bemba-ASR-Corpus
Citation (if available): N/A
Data set developer(s): Claytone Sikasote, Antonis Anastasopoulos
Data statement author(s): Claytone Sikasote, Anastasopoulos
Others who contributed to this document: N/A
Link to the dataset: https://github.com/csikasote/bemba-language-corpus
Dataset license: Yet to set!
Total Number of hours: 24 hours

A. CURATION RATIONALE

The BembaSpeech is an automatic speech recognition(ASR) dataset for Bemba language based on speech utterances recorded from text obtained mostly from Bemba literature books. The dataset has over 14, 000 utterances culminating into approximately 24hours of speech data.

The motivation for building the BembaSpeech dataset is to create a speech recognition dataset for Bemba language that can be used to train and test speech recognition systems.

B. LANGUAGE VARIETY/VARIETIES

The language considered in this corpus is Bemba (ISO 639-3 bem). Bemba is a bantu languages native to north-eastern Zambia. The language is also spoken in some parts of Democratic Republic of Congo, Tanzania as well as Botswana.

C. SPEAKER DEMOGRAPHIC

Speakers were directly approached to create audio utterances by eliciting text scripts in the Lig-Aikuma mobile application. The speakers were selected based on their fluency to speak and read Bemba and not necessarily native language speakers. It is, however, expected that some, but not all, of the speakers speak Bemba as a native language. Based on information extracted from metadata as supplied by speakers on recording, seventeen (17) speakers were involved in recording audio, of which eight (8) are female and nine (9) male. The age range of speakers is between 22 and 28 years and, all of them are identified to be black. In terms of occupation, all the speakers are university students.

D. ANNOTATOR DEMOGRAPHIC

No annotations were done in this datasets, therefore, no annotator demographic informations is available.

E. SPEECH SITUATION

The corpus comprises of the speech utterances that were recorded from scripts using the Lig-Aikuma application using the text elicitation mode. The recordings were done in Zambia at the University of Zambia between 20th August, 2020 and 12th October, 2020.

F. TEXT CHARACTERISTICS

Most of the text in the dataset is from private and publicly available Bemba literature books. Some of it was obtained from transcriptions of movies and recorded TV/radio programs. A small portion of approximately 5% was obtained from a Christian magazine.

G. RECORDING QUALITY:

The audio utterances were recorded using the Lig-Aikuma mobile application by eliciting texts that are tokenized at sentence level. Almost all recordings were not done in closed and soundproof environment. Therefore, It is expected that there would be some background noise.

H. OTHER

Some of the texts contained in this dataset have been used without permission from copyright owners. However, permission has been sought awaiting to be granted.

I. PROVENANCE APPENDIX

Nothing contained in this dataset is obtained from an already existing dataset. Therefore, provenance appendix is not available.