BK3AT: Bangsamoro K-3 Children's Speech Corpus for Developing Assessment tools in the Bangsamoro Languages

Kiel Gonzales, Jazzmin Maranan, Nissan Macale, Edsel Jedd Renovalles
Nicole Anne Palafox, Francis Paolo Santelices, Jose Marie Mendoza

University of the Philippines Diliman, Philippines
{gonzales.kiel, jazzreyesmar, nissan.macale,e.jedd.mr, nicoleaapb}@gmail.com
{francis.santelices, jose.marie.mendoza}@eee.upd.edu.ph

Abstract

Bangsamoro languages are among the under-resourced languages in the Mindanao region in the Philippines. Moreover, there is no currently publicly available data for children's speech on most of these languages. BK3AT children's speech corpus is a corpus designed for creating speech technologies that could help facilitators and teachers in K-3 education. The corpus consists of 122 hours of children speech data across 10 languages: Bahasa Sug, Chavacano, English, Filipino, Iranun, Maguindanaon, Meranaw, Sinama, Teduray, and Yakan. Preliminary experiments using Wav2Vec-XLSR architecture have been done in fine-tuning the Tagalog and L2 English corpus subsets to develop automatic speech recognition backend for literacy assessment. Results from the experiments show low word error rates (WERs) for small-vocabulary and targeted domains.

Keywords: children's speech corpora, low-resource languages, Bangsamoro languages

1. Introduction

The Bangsamoro Autonomous Region in Muslim Mindanao (BARMM) is home to at least 4 million Filipinos of distinct and diverse indigenous and Islamic cultures (Philippine Statistics Authority, a). They are using at least 13 languages including Filipino, Arabic, English, Cebuano, Sabah Malay, Meranaw (Maranao), Yakan, Bahasa Sug (Tausug), Sinama (Sama), Iranun, Chavacano, Teduray (Tiruray), and Maguindanaon. From among these languages, only Tagalog, Cebuano, and Maguindanaon are in the top ten leading languages used at home according to the census of the Philippine Statistics Authority (Philippine Statistics Authority, b). The available speech corpora on languages used in BARMM would be little to none especially with children's speech data.

In 2022, The Bangsamoro K-3 Assessment Tools (BK3AT) Project was launched through the funding of the Australian government through Education Pathways to Peace in Mindanao (Pathways), in partnership with the Department of Education (DepEd) and the Ministry of Basic, Higher, Technical Education (MBHTE) and Readability Center. The objective of the project is to develop an assessment tool kit that will provide the educators and eventually to policymakers information on the performance of the Bangsamoro K-3 students in the domains of numeracy, literacy, and social emotional learning.

The automated literacy assessment of the tool kit requires the development of an automatic speech recognition (ASR) and language modeling. Hence, the need for the creation of a Bangsamoro Children's speech corpus. Not only can the corpus be used for developing assessment tools, but also for other applications like phonological awareness and reading tutors.

2. Data Design and Collection

Developing ASR systems requires data relevant to your application. It is important to obtain clean and accurate speech utterances in order to have a usable ASR, at the least. This section details the process of collecting children's speech data including the tools used and setup.

2.1. Design

The BK3AT Children's corpus was designed to be the baseline data which the software developers and engineers can use as models for the literacy assessment. It consists of 10 languages: Filipino, English, and 8 mother tongue languages used in BARMM namely: Bahasa Sug, Chavacano, Iranun, Maguindanaon, Meranaw, Sinama, Teduray, and Yakan. The prompts for every language consists of four different types of texts: words, phrases, sentences, and passages. The prompts were first created in Filipino and were listed in increasing difficulty. Then the seed prompts were shared with translators recommended by the MBHTE to create a similar corpus. The mother tongue language prompts are not translated word for word but rather follow the structures of the syllables and the increasing difficulty as in the Filipino prompts. In addition to the structure, the corpus should cover all the phonemes of the language and the texts should be at level or age appropriate for Grade 2 and Grade 3 students.

The requested participants for the data collection are Grade 1-3 students coming from all divisions of BARMM. They are comprised of instructional or independent readers in order to gather correctly read prompts over recordings containing miscues. They were asked to read three languages: Filipino, English, and their mother tongue language. In addition to the three languages, the participants were also requested to read English letters.

2.2. Data Consent

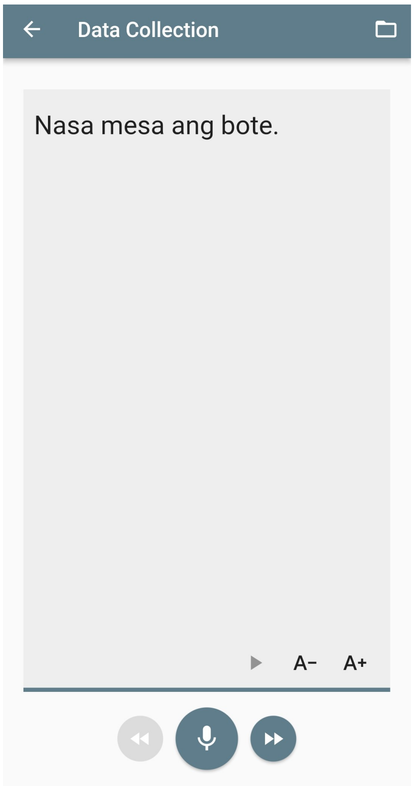
To protect of the identity of the participants, a data consent form was given to the parents of the participants through their class advisers to request for their permission to be recorded. The data consent form contains the description of the project and the recording activity. The parents are informed that the participants will be asked to read a set of prompts and have their voice recorded in three languages. In addition to the asking for permission for the audio recording, taking of pictures for documentation was also included in the consent form. Only those participants with signed parent consent forms are included in the activity.

The time slots per participant per language is at 30 minutes each. If they are not able to finish on time, the recording will be stopped and not force the participant to finish all the remaining prompts. They can also request for a break should they need to rest. Moreover, the participant is free to back out from the session anytime and the session will not be included in the corpus.

The names of the participants were redacted in the speech corpus. Only the information on age, gender, and mother tongue language will be included. Furthermore, their identity is kept confidential in reports by not mentioning the names and blurring the face of the participants in the photos taken.

2.3. Recording Tool

An audio recording software was used to facilitate the collection of speech data. However, data collection in BARMM involved addressing some limitations. These limitations include not having computers on hand, unstable internet connections, and not having the proper recording equipment required for a clean recording. Since android phones are more accessible than computers in BARMM, a recording tool that is compatible with Android devices (RecTool Mobile) was developed using the Flutter[1] framework. It is an application that is specific for collecting speech utterances which has a simple user interface as shown in Figure 1.



Nasa mesa ang bote.

Figure 1: BK3AT RecTool Mobile interface

For each recording session, the speaker is presented with the prompts to be read. The selection and order of prompts is done automatically by the recording tool. After pressing the record button, the speaker starts to read the prompt which could be a word, a phrase, a sentence or a passage. The facilitator ensures that the speaker completes reading text before pressing the stop button to proceed to the next prompt. The recording tool is operated by a volunteer teacher in BARMM.

The recording tool is also used to collect information about the speaker. This information includes the speaker's age, gender, profession, first language and the first languages of the speaker's parents. The information about the first language is further differentiated by adding the region where the speaker or speaker's parents grew up, which is how we approximate the dialect spoken. The collected information is used to categorize the speakers and easily monitor the distribution of speakers per language according to age, gender and dialect.

2.4. Recording Setup

The data collection was done through the assistance of teachers in BARMM. They were given an online orientation by the BK3AT tech team so they are aware of the prescribed recording set-up and the proper usage of the recording tool. A recording

[1] https://docs.flutter.dev/

kit shown in Figure 2 which consists of a headset, a splitter, earphones, and a flash drive were shipped to the data collectors for a consistent hardware setup. The prompts to be recorded, along with the installer for RecTool Mobile were stored in the USB OTG flash drives.

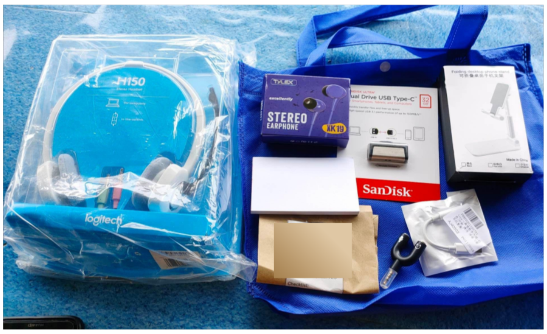


Figure 2: Equipment used for BK3AT children's speech corpus data collection.

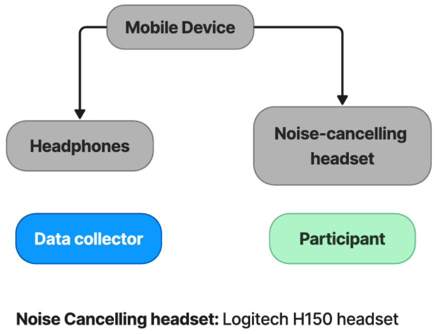
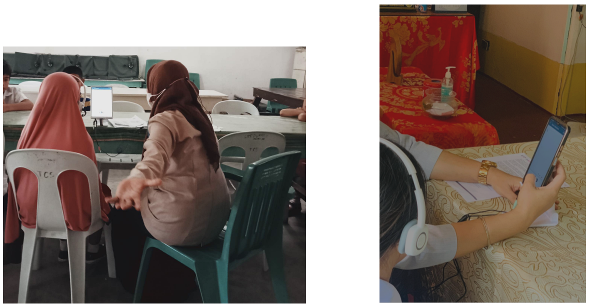


Figure 3: Diagram of the recording setup

A noise reduced headset was provided to the child to be able to concentrate during the recording session. On the other hand, the teacher also used earphones to properly hear the utterance. If mistakes are heard, the participant is asked to repeat the recording of the prompt. This setup is shown in Figure 3

The recordings were mostly done in empty classrooms or admin offices to minimize the noise. Figure 4 shows two examples of the setup. The participants are given a 30-minute time slot per language to provide an ample time to complete the recording. The data collectors then uploaded the recorded audio files on an online sharing folder for accessibility. The files are organized in a structure illustrated in Figure 5 where directories of languages contains the data of each speaker. Specifically, each utterance of the speaker is matched with its ground truth transcription which are the prompts presented during recording. All of these are compiled in a *.log* file together with the speaker's metadata and session ID.



(a) Classroom recording setup

(b) Small room recording setup

Figure 4: Data collection recording setup for BK3AT children's speech corpus

3. Corpora Details

3.1. Corpora Statistics

Summary statistics for the BK3AT Children's speech corpus are shown in Table 1. The corpus details are divided per language. The BK3AT Children's speech corpus currently contains 130,733 recordings from over 244 speakers of 10 different BARMM languages. This corresponds to over 122 hours of recorded read speech. A language corpus in the BK3AT Corpora has at least 4 hours of recording (Maguindanaon) to 45 hours (Filipino). The combined recording prompts used for data collection correspond to 352,785 tokens, where a token can be a word, number, acronym etc. used in the text.

In the data collection for each language, the majority of participants are female, compromising a percentage of the total speakers ranging from 57.14% for Bahasa Sug (20 female and 15 male), up to 76.67% for Iranun, Sinama, and Yakan (23 female and 7 male). The only exception is Teduray, where the majority of the speakers are male (14 female and 16 male). It is noteworthy that genders were not recorded for some participants in English and Filipino (6.15% and 6.61% of their populations, respectively). We also examined the age distribution of our speakers per language, and histograms of the speaker ages are shown in Figure 6. The means of speaker ages range fr‘om 9 for Maguindanaon and Yakan to 12 for Teduray. Meanwhile, the highest standard deviation of ages was reported at 2.93 for Filipino.

3.2. Licensing and Availability

The BK3AT Children's speech corpus is owned by Department of Foreign Affairs and Trade (DFAT) Australia and Ministry of Basic, Higher and Technical Education. Access to the dataset can be requested to the aforementioned agencies. Upon creation, it is licensed under Creative Commons

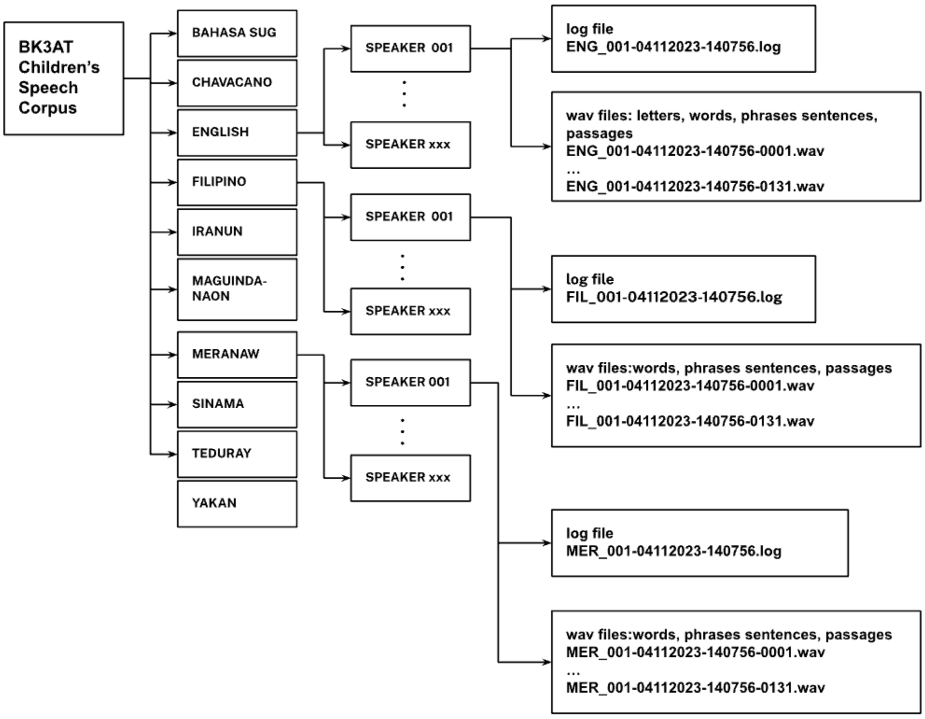

BK3AT Children's Speech Corpus
BAHASA SUG
CHAVACANO
ENGLISH
FILIPINO
IRANUN
MAGUINDA-NAON
MERANAW
SINAMA
TEDURAY
YAKAN
SPEAKER 001
SPEAKER xxx
SPEAKER 001
SPEAKER xxx
SPEAKER 001
SPEAKER xxx
log file
ENG_001-04112023-140756.log
wav files: letters, words, phrases sentences, passages
ENG_001-04112023-140756-0001.wav
...
ENG_001-04112023-140756-0131.wav
log file
FIL_001-04112023-140756.log
wav files:words, phrases sentences, passages
FIL_001-04112023-140756-0001.wav
...
FIL_001-04112023-140756-0131.wav
log file
MER_001-04112023-140756.log
wav files:words, phrases sentences, passages
MER_001-04112023-140756-0001.wav
...
MER_001-04112023-140756-0131.wav

Figure 5: BK3AT Children's Speech Corpus Structure

Attribution-Non-Commercial (CC-by-NC 4.0).

4. Corpora Use

4.1. Speech-to-Text Systems

The English and Filipino subset of the BK3AT Corpora were used to develop children speech recognizers (CSRs) integrated in the Bangsamoro K-3 Assessment Tool (BK3AT) in order to detect reading miscues and evaluate the Bangsamoro K-3 students' phonological awareness and reading skills. The use of ASRs to aid the assessment of students' literacy have been implemented in other research such as an automated reading tutor [(Pascual and Guevara, 2012)].

The systems were implemented using Wav2Vec2 [Baevski et al. (2020)], a self-supervised speech system. Specifically, the CSR model was built using XLSR-Wav2Vec2 [Grosman (2021)] a pre-trained speech model, which performs at a Word Error Rate (WER) of 7.33% tested on the Common Voice 11.0 Corpora. The aforementioned model was tested on 5.86 hours of multi-speaker data from the English BK3AT subset, achieving a WER of 47.49%. To further improve the recognition of the model, a language model (LM) was incorporated. The KenLM Language Model Toolkit [Heafield (2011)] was used to create a language model for the English BK3AT prompts. By incorporating an LM boost to the model, the recognition of the same test data improved to a WER of 33.31%.

For the BK3AT Filipino subset, a similar approach was explored. An English-Filipino speech topic tagger [Tumpalan and Recario (2023)] with the same model but trained on an open-source Filipino dataset [MagicHub (2022)] resulted in 26.8% WER. This model was used as a baseline for the Filipino CSR model.

The proposed Filipino CSR model yielded unrecognizable results or a WER of 100% when the system was evaluated solely using Jonatas' XLSR-Wav2Vec2 model, thus it was further fine-tuned on the BK3AT Filipino subset using 0.236 hours of data. Learning rate of 0.0003 was used for fine-tuning. Training ran for a maximum of 300 steps with a batch size of 1 while evaluation ran for 200 steps with a batch size of 2.

The fine-tuned model was then tested using 14.71 hours of the Filipino subset, achieving a WER of 61.66%. Similar to the English CSR model, a LM boost was implemented to improve the recognition, acheiving a 50.59% WER.

Table 2 summarizes the fine-tuned data, test data, and WER performances of the English and Filipino CSR models.

5. Future Work

Currently, the developers are still working on improving the assessment tool including the fine-tuned backend ASR previously mentioned. Meth-

Language	Gender	Speaker Count	Utterance Count	Total Audio Duration (h:m:s)	Tokens	
					Total	Unique
Bahasa Sug	F	20	4,107	04:48:06	10,650	217
	M	15	3,081	03:48:26	7,991	217
	all	35	7,188	08:36:32	18,641	217
Chavacano	F	19	4,073	03:44:50	11,919	132
	M	11	2,199	01:50:18	6,429	132
	all	30	6,272	05:35:08	18,348	132
English	F	169	22,038	16:13:49	63,072	155
	M	75	9,780	07:44:06	27,959	155
	all	244	31,818	23:57:55	91,031	155
Filipino	F	157	29,208	28:47:24	84,404	212
	M	83	15,427	16:58:42	44,554	212
	all	240	44,635	45:46:06	128,959	212
Iranun	F	23	4,630	05:16:00	13,546	227
	M	7	1,446	01:58:52	4,257	227
	all	30	6,076	07:14:52	17,803	227
Maguindanaon	F	20	3,459	02:52:41	7,677	183
	M	10	1,732	01:22:51	3,831	183
	all	30	5,191	04:15:33	11,508	183
Meranaw	F	21	4,432	04:34:48	13,299	210
	M	9	1,943	02:10:05	5,882	210
	all	30	6,375	06:44:53	19,181	210
Sinama	F	23	3,514	03:45:28	7,901	167
	M	7	1,069	01:17:21	2,404	167
	all	30	4,583	05:02:50	10,305	167
Teduray	F	14	2,937	03:10:36	7,535	263
	M	16	3,331	03:41:28	8,537	263
	all	30	6,268	06:52:04	16,072	263
Yakan	F	23	9,451	06:15:47	16,048	291
	M	7	2,876	01:53:42	4,889	291
	all	30	12,327	08:09:30	20,937	291
Total	**-**	**244**	**130,733**	**122:15:23**	**352,785**	

Table 1: Summary statistics for the BK3AT Corpora.

Language	Total Audio Duration	Duration of Fine-tuned Data	Duration of Test Data	Word Error Rate (WER)	
				w/o LM	w/ LM
English	∼24 hours	-	5.86 hours	47.49%	33.31%
Filipino	∼45 hours	0.236 hours	14.71 hours	61.66%	50.69%

Table 2: Summary of the fine-tuned and test data durations and the WER performances of the English and Filipino CSR models.

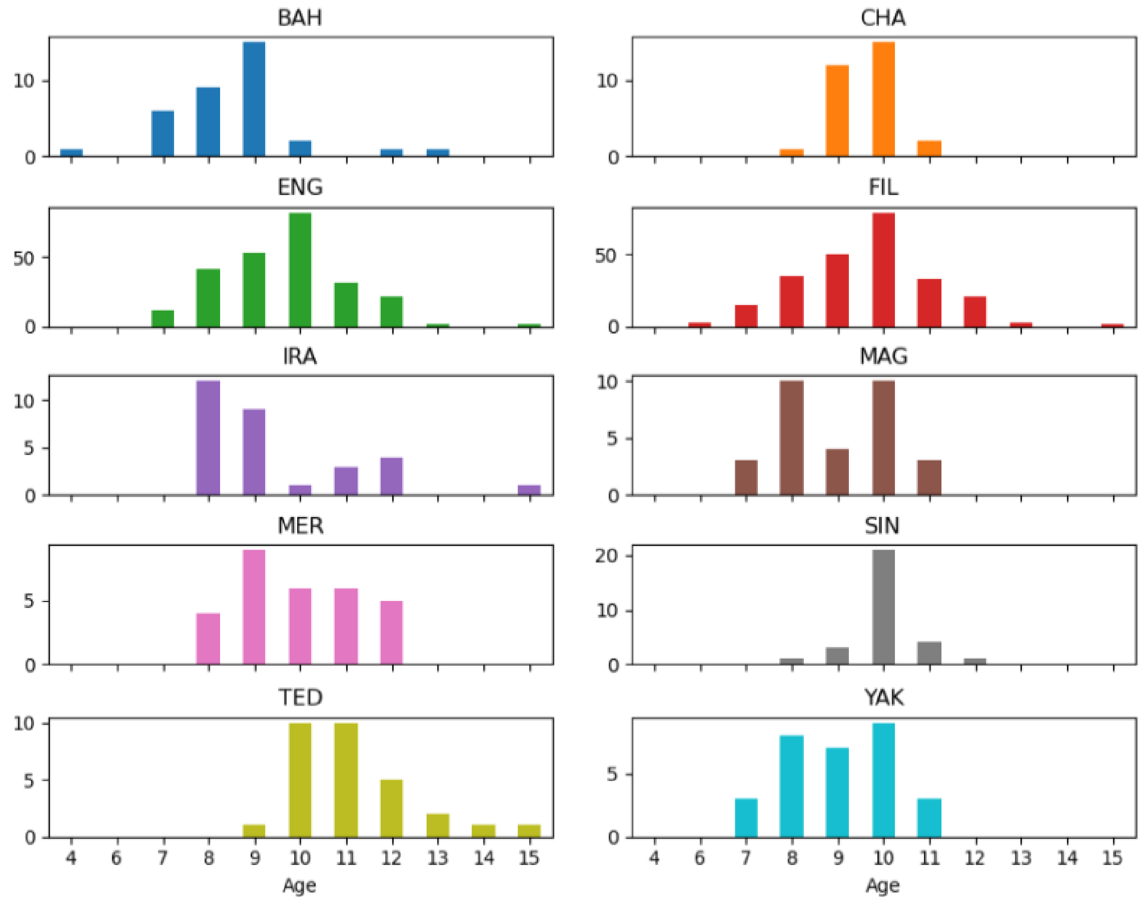


Figure 6: Speaker age distribution of different languages in BK3AT children's speech corpus
Bahasa Sug (BAH), Chavacano (CHA), English (ENG), Filipino (FIL), Iranun (IRA), Maguindanaon (MAG), Meranaw (MER), Sinama (SIN), Teduray (TED), Yakan (YAK)

ods such as language model (LM) boosting, pretraining, and data augmentation are being explored and implemented to further utilize the corpus for its intended application. For future work, the team envisions completion of automated literacy assessment for all the BARMM mother tongue languages.

6. Acknowledgements

We would like to thank the Ministry of Basic, Higher and Technical Education (MBHTE) of Bangsamoro Autonomous Region in Muslim Mindanao (BARMM), Education Pathways to Peace in Mindanao (Pathways) with special mention to Ma'am Bonna Duron Luder, Department of Foreign Affairs and Trade (DFAT) Australia, Readability Center led by Jaycee Pascual. Special thanks to Michael Gringo Bayona and Crisron Rudolf Lucas for helping out in this paper.

7. Bibliographical References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Jonatas Grosman. 2021. Fine-tuned XLSR-53 large model for speech recognition in English. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

MagicHub. 2022. ASR-SFDUSC: A scripted filipino daily-use speech corpus. https://magichub.com/datasets/filipino-scripted-speech-corpus-daily-use-sentence.

Ronald Pascual and Leidy Guevara. 2012. Developing an automated reading tutor in filipino for primary students.

Philippine Statistics Authority. a. Highlights of the Philippine Population 2020 Census of Population and Housing (2020 CPH).

Philippine Statistics Authority. b. Tagalog is the Most Widely Spoken Language at Home (2020 Census of Population and Housing).

John Karl B. Tumpalan and Reginald Neil C. Recario. 2023. English-filipino speech topic tagger using automatic speech recognition modeling and topic modeling. In *Advances in Information and Communication*. Springer Nature Switzerland.