Breaking the Low-Resource Barrier for Dagbani ASR: From Data Collection to ASR Modeling

Anonymous authors
Paper under double-blind review

Abstract

This paper presents a data collection pipeline and process for a transcribed Dagbani audio dataset. Dagbani is an African language spoken predominantly in Ghana and in parts of northern Togo. We apply the data to build the world's first automatic speech recognition (ASR) system for Dagbani. We hope this methodology can serve as a blueprint or guideline for other similar efforts.

1 Introduction

Languages of Northern Ghana include Dagaare, Dagbani, Gonja, Buli and Gurune (Azunre et al., 2021), to mention just a few. These languages are even lower resourced than southern Ghanaian languages such as Twi. This is likely due to the region being further away from the historical commercial centres along the southern coast, leading to lower levels of local language digitisation, education and resources, as well as higher levels of poverty (Yaro & Hasselberg, 2010). These languages are yet to evolve from simple (or in some cases very little) online existence to optimal online presence (Azunre et al., 2021) (Tiayon, 2013).

A couple of mostly self-funded grassroots projects have attempted to remedy this, notably GhanaNLP and Dagbani Wikimedia Group. For example, GhanaNLP has deployed a neural machine translation system for Dagbani. The Language Identification work presented in Adebara et al. (2022) incorporates all Northern Ghanaian languages identified by Azunre et al. (2021) into its framework. On the speech front, however, no transcribed Dagbani speech data is available. Thus, no ASR models existed - until now.

In low-resourced settings such as this, transcription of audio often requires significant human capital (Dar'gis et al, 2020). In this paper, we highlight a data collection pipeline that streamlines this process and enables collection of sufficient data for ASR models at low cost. Considering the limited availability of speech datasets for most African languages (Ritchie et al., 2022), we hope this can serve as a blueprint for other similar efforts.

The tools and methodology emphasise free and unrestricted access to data, with an additional goal of easing generalisation to other languages and domains.

In this paper, we present a transcribed Dagbani speech dataset and ASR system developed for Dagbani from the resulting data. This is the first ever ASR system for the language, to the best of our knowledge. The trained ASR model is available via [REDACTED FOR BLIND REVIEW]. The resulting data from this undertaking is hosted on Wikimedia Commons, allowing free unrestricted access to all individuals, via [REDACTED FOR BLIND REVIEW].

2 Focus language

In this initial iteration of the data collection pipeline, we focus on Dagbani. We collect some Dagbani sentences and corresponding audio pairs, and proceed to build an ASR system with the resulting dataset. The insights drawn from this project, however, have the potential to serve as a blueprint or guide for similar efforts in both Dagbani and other low-resourced Ghanaian languages, particularly the Northern Ghanaian languages such as Dagaare, Wali, Mampurili and Gurune (Frafra).

2.1 Overview

Dagbani is a Gur language spoken by over 3 million people primarily across the northern region of Ghana and sparsely in northern Togo (Eberhard et al., 2022). It is a compulsory language in primary and junior high schools in the country (Eberhard et al., 2022) and serves as a principal language for many day-to-day transactions. Despite its widespread use, however, Dagbani remains a low-resourced language, with limited availability of data for research and development of language technologies.

2.2 Orthography

Initial Dagbani writing system was based on Ajami, the modified version of the Arabic script used in writing many African languages (Ajura, 1959). Currently the Dagbani writing system is based on the Latin script, including the apostrophe and the letters ɛ, ŋ, ɔ, ɣ and ʒ, and the digraphs ch, gb, kp, ŋm and ny (Olawsky, 2003). The orthography currently used (Eberhard et al., 2022) (Baldi & Mahmoud, 2006) represents a number of allophonic distinctions, but does not feature any diacritic marks.

2.3 Tone

Gur languages feature a verbal aspect marking. Like most languages in this family, Dagbani is a tonal language that employs pitch to distinguish between words, notable examples being gballi [ɡbálːɪ́] (high-high) and gballi [ɡbálːɪ̀] (high-low) (Olawsky, 1997) – wherein the former represents a “grave”, the latter represents a “zanamat” (a mat of woven grass). Particularly, the tone system of Dagbani features two level tones and downstep – a phenomenon in tonal languages in which if two syllables have the same tone, the second syllable takes on a lower pitch than the first.

In writing, however, tone is not marked (Eberhard et al., 2022)(Baldi & Mahmoud, 2006). As such, it is not always clear why a word is written in a certain way (Olawsky, 1995).

3 Dataset creation

Similar to Ogayo et al. (2021), the project has at its core an open participatory philosophy, allowing all interested parties to contribute their voices for Dagbani.

Additionally, we have adopted a Creative Commons Zero (CC0) licence for all voices built, allowing individuals free access to the dataset. The models we developed on this data are available via an API for easy deployment of applications.

3.1 Developing text data for reading

We first collect textual data for Dagbani by extracting sentences from articles posted on the Dagbani Wikipedia page[1]. These articles are governed by the CC0 licence, allowing their use for any purpose. All of the data was written and published in the target language. Sources from which these articles were developed include folklore and legends, articles about prominent individuals written from scratch and articles obtained by translating English Wikipedia articles into Dagbani.

Despite the Dagbani Orthography Committee (1998) developing a standard orthography for the language, there has not been widespread adoption of the system, resulting in several instances of alternate spellings for the same word, for example yɛltɔɣa and yɛltoɣa. Also, the textual data extracted from the Dagbani Wikipedia pages comprised a blend of sentences written in the Eastern Dagbani dialect and others in the Western Dagbani dialect. This, however, is not a considerable limiting factor, considering that the dialects are mutually intelligible and only differ slightly in the root vowels of certain lexemes.

3.2 Voice talent recruitment

The recruitment of voice talent for the Dagbani speech data collection project was conducted following the criteria of fluency, literacy, and familiarity with voice recording as outlined in Niekerk et al. (2017). A total of 24 participants were recruited from the Tamale area, who were fluent and literate in Dagbani and had prior experience in data collection. Participants were remunerated with cash for their participation in the project.

3.3 Audio recording

Due to logistical constraints, it was impossible to record audio in a studio environment, which would have been preferable. As such, the current version of the Dagbani speech dataset was recorded in residential homes with smartphones and ear phones/AirPods to reduce as much background noise as possible and to ensure clarity of utterances.

Participants on the Android system used the Spell4Wiki app[2], an android application for making voice recordings of entities and uploading to Wikimedia commons. All audio data recorded with this tool were saved in the '.ogg' format and thus had to be converted to '.wav' format during the data aggregation process. The files were named with the content of the sentence (where spaces were replaced with '_') to make aligning transcriptions with corresponding audio recording relatively easy. The one drawback with this method, however, is that Spell4Wiki limits all input text to a maximum of 30 characters, making participants work with sentences of shorter lengths.

Unfortunately, Spell4Wiki only runs on the Android platform. As such, voice talent on the iOS system extracted sentences into Google Sheets and used a voice recorder app to record the corresponding audio for each sentence in the '.wav' format. These were then uploaded to a Google Drive folder set up for the project. Participants using this method were able to record audio of longer lengths, since there's no limit on the number of characters that can be entered into a cell in Google Sheets.

Overall, the entire data collection process – from extracting sentences to recording audio – spanned three weeks, allowing participants ample time to iterate and produce clearer utterances.

[1] https://dag.wikipedia.org/wiki/Di%C5%8B%27gahim:AllPages

[2] Spell4Wiki mobile app available at: https://github.com/manimaran96/Spell4Wiki

Table 1: Dagbani Speech Corpus Metadata Overview

ATTRIBUTE	VALUE
MALE	15
FEMALE	7
UTTERANCES	11,207
UNIQUE TOKENS	7,222
TOTAL TOKENS	48,636
TOTAL DURATION	9hr 34 min

3.4 ALIGNING AUDIO TO TRANSCRIPTIONS

Spell4Wiki uses the content of the textual input (wherein spaces are replaced with '_') as identifiers for uploaded audio files. Thus, to align audio recordings to the appropriate transcript after downloading from wikimedia commons, we strip the extension part from the filename and replace the '_'s with spaces.

For iOS users who used Google Sheets and Google Drive for the data collection process, we devised a naming convention of the form 'username-#' for the audio data, where 'username' represents the participants user id and '#' represents the sentence number (i.e. the row number of the cell holding the textual data in the Google Sheet). This convention allowed us to easily align the data on the sheet with the audio files in google drive.

For both methods of data collection, we enlisted the expertise of 5 proof readers to check that all audio files and transcription pairs matched.

3.5 VALIDATION AND QUALITY CONTROL

For languages with multiple dialects, the traditional approach would be to remove sentences with mixed dialects or replace them with an equivalent sentence with one dialect. We however thought it best not to do so for Dagbani as the dialects are mutually intelligible and utterances with mixed-dialects do not really show any marked differences. As such doing so would not result in any appreciative gain in dataset quality.

Everyday speech for Dagbani, like most African languages, contains borrowed words from other languages, predominantly English. Removing such borrowed words from the dataset would not be ideal, as doing so would result in a dataset that was not truly representative of all topics and concepts in normal day-to-day discourse. Hence, we opted to leave such words in the exact form they take in their original languages, but to have participants employ the pronunciation adopted for such words by speakers of Dagbani, seeing as such pronunciations are most natural to the target audience and most often than not, do not really differ from the actual pronunciation in the original language (i.e. language from which they were borrowed).

Overall, recorded audio was validated on many fronts to ensure that there wasn't any destructive noise in the final product, and to ensure that the transcripts did not perpetuate any prejudices and biases. This makes the data suitable for language modelling applications for Dagbani in future works.

3.6 Dataset license

All data items are governed by the Creative Commons Zero (CC0) License, allowing all individuals unrestricted access to the dataset, and granting permissions to individuals to use the data for any purpose.

Permissions were sought from all 22 participants in the following capacities:

- to use participant's voice to train ASR and possibly TTS systems.
- for participant's voices to be made freely available to all individuals and organisations.

4 ASR system

We built a Speech-to-Text (or ASR) system based on the Wav2Vec 2.0 architecture introduced in Baevski et al. (2020) with the newly created dataset. Using Hugging Face Transformers Python library, the wav2vec 2.0 checkpoint was fine-tuned on the data presented in this paper. 95% of the data was used for training and 5% was used for validation. Convergence was achieved in 9 epochs and about 2 hrs, on a Dell XPS Windows Desktop with an 8GB NVIDIA GeForce RTX 2070 SUPER GPU. The word-error-rate (WER) metric achieved by our model was 0.341. We know from practical experience that this level of performance is sufficient for noncritical applications, and this was confirmed in testing. In fact, we observed that in most instances where the model outputs a wrong transcript, the difference from the ground truth is not significant, and can be often be accounted for via alternate spellings of the same word. This stems from the aforementioned lack of general consensus on the orthography for the language, with multiple spellings emerging for many words in Dagbani. In other words, the practical WER value perceived by users is significantly higher than 0.341. The trained ASR model is available via [REDACTED FOR BLIND REVIEW].

5 Challenges and limitations

The methodology we employed for the data collection process was not without its fair share of limitations. Particularly, the 30-character limit built into the Spell4Wiki app meant that participants who used that system had to record shorter sentences.

Given the potential of the application for creating data for low-resourced languages, we are currently in talks with the Spell4Wiki developers to extend the character limit and duration of recordings significantly. However, iPhone users were able to record sentences of longer lengths as Google Sheets and the recorder app used had no such limits. Overall, our data has a nice representation of typical utterances an average user might reference within the app.

Additionally, we do not make any distinctions between the Eastern and Western Dagbani dialects, nor do we perform any prosodic annotation of the dataset and therefore do not benefit from any boost in quality and/or performance such annotations may present.

6 Ethical considerations

We address the following ethical considerations to ensure that the entire data collection process is respectful, culturally sensitive and benefits the language and research communities.

- Informed consent: we seek for and obtain consent from all individuals who participated in the project.

- Privacy and confidentiality: given the open nature of the dataset created, we only provide recordings and transcriptions that do not breach the privacy of any individual or organisational entity.
- Cultural sensitivity: we employ rigorous screening to ensure that voice talents do not annotate any sentences that perpetuate cultural bias, and to avoid practices that might be considered disrespectful, taking into account local customs and traditions.
- Benefit sharing: we allow free unrestricted access to the dataset. Also, we have integrated the ASR model into the Khaya app for use by all individuals.

7 Conclusion

In this paper, we describe creating speech data for the low-resourced African language Dagbani from scratch and with a limited budget. We demonstrate how the outcome can enable researchers and developers to build Automatic Speech Recognition (ASR) systems with the resulting data. Particularly, through this paper, we open the gateway to building similar datasets and ASR systems for other Northern Ghanaian Languages such as Wali, Mampurili, Gurune (Frafra), Buli, Gonja, Dagaare and many more (Azunre et al., 2021) (Tony, 1989) (Bendor-Samuel, 1989).

We present the data collection pipeline employing Spell4Wiki and Wikimedia Commons for creating truly open and freely accessible speech data for endangered languages. In the not-so-distant future, we hope to expand the geographical coverage of the project and resulting datasets, and to produce data that will be suitable for building both ASR and TTS systems. Last but not least, future work should look into providing prosodic contexts for utterances so that the dataset produced might harbour the potential to be employed on a wider variety of tasks.

○ Author Contributions

To be completed post-review, to maintain anonymity

○ Acknowledgments

To be completed post-review, to maintain anonymity

References

Ife Adebara, Abdel Rahim Elmadany, Muhammad Abdul-Mageed, Alcides Alcoba Inciarte. AfroLID: A Neural Language Identification Tool for African Languages. arXiv:2210.11744v3 [cs.CL] 7 Dec 2022

Afa Yusif Ajura (1959) Dagbanli Ajami and Arabic Manuscripts of Northern Ghana [Available online at http://hdl.handle.net/2144/32937]

Richard Asante and E. Gyimah-Boadi (2004), Ethnic Structure, Inequality and Governance of the Public Sector in Ghana. United Nations Research Institute for Social Development.

Paul Azunre, Salomey Osei, Salomey Afua Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabanka, Bernard Opoku, Clara Adare Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Issac K.E. Ampomah, Joseph Otoo, Reindorf Borkor, Sandylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. (2021) NLP for Ghanaian Languages: arXiv:2103.15475v2 [cs.CL]

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli. (2020) Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations [arXiv: 2006.11477]

Sergio Baldi, Adam Mahmoud (2006). Dagbani basic and cultural vocabulary. Unive. degli studi di Napoli “L'Orientale''. p. 10

John T. Bendor-Samuel (1989). The Niger-Congo Languages. Lanham, MD: University Press of America.

Dagbani Orthography Committee. (1998). Approved Dagbani Orthography. n/p (Tamale, N/R)

R. Dar‘gis, P. Paikens, N. Gruzitis, I. Auzina, and A. Akmane, “Development and evaluation of speech synthesis corpora for Latvian,” (2020) in Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, May 2020, pp. 6633–6637. [Online]. Available: Development and Evaluation of Speech Synthesis Corpora for Latvian - ACL Anthology

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). (2022) Ethnologue: Languages of the World. Twenty-fifth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Tony Naden (1989). Gur. Lanham, MD: University Press of America. pp. 141–168.

D. van Niekerk, C. van Heerden, M. Davel, N. Kleynhans, O. Kjartansson, M. Jansche, and L. Ha. (2017) “Rapid development of TTS corpora for four South African languages,” in Proc. Interspeech 2017, pp. 2178–2182. [Online]

Perez Ogayo, Graham Nubig, Alan W. Black. (2021) Building African Voices: arxiv: 2207.00688v1

Knut Olawsky. (1995) Application for an orthographical survey of Dagbani. Unpublished ms. Düsseldorf.

Knut Olawsky (1997). Interaction of tone and morphology in Dagbani. Online version: https://user.phil.hhu.de/~olawsky/hp-dgfs.htm

Knut Olawsky (2003) “What is a word in Dagbani”, Word, Cambridge University Press, pp. 205-226

Sandy Ritchie, You-Chi Chieng, Mingqing Chen, Rajiv Mathews, Daan van Esch, Bo Li, Khe Chai Sim: Large vocabulary speech recognition for languages of Africa: multilingual modeling and self-supervised learning. arXiv:2208.03067v2 [cs.CLI] 4 Oct 2022

Charles Tiayon, 2013. About professional writing and translation in African languages. [Available at https://metaglossia21.blogspot.com/2013/03/about-professional-writing-and.html?m=1]

Joseph A. Yaro and Jan Hasselberg, 2010. The contours of poverty in northern Ghana: policy implications for combating food insecurity. Institute of African Studies Research Review, 26(1):81-112