

GhanaNLP

PUTTING GHANAIAN LANGUAGES ON THE GLOBAL
MACHINE LEARNING MAP

Dr. Stephen Edward Moore



GhanaNLP

Open Source Movement of like-minded volunteers

Time and skills dedicated to building an ecosystem of :

1. Open-source datasets
2. Open source computational methods
3. An army of NLP researchers, scientists and practitioners

To revolutionize and improve every aspect of Ghanaian life through the powerful tool of Natural Language Processing.

Sponsors and Partners



Microsoft



Dataset Collection

- ❑ We set out a **google document** and requested for English to Twi translation.
- ❑ We used **Jehovah Witness (JW300)** translated English to Twi articles.
- ❑ We also used the **Christian Bible** as a main source of translation.
- ❑ Some **English to Twi books** on the market for tourists.

Challenges and Solutions

- ❑ Data contains non-twi words.
 - ❑ E.g. JW300 had over 600K sentences.
 - ❑ Several non-twi words.
 - ❑ Finally ~22K uniques words
- ❑ Involving **Ghana Linguistic Association**
- ❑ University Researchers and Professors reviewing some sentences.
- ❑ Several translations of same sentences from **google document**.

Dataset Sharing

- ❑ All our datasets are **open source**.
- ❑ Most data collected is volunteered.
- ❑ Other linguist and researchers are willing to share their data.
- ❑ Exchange is no robbery!
- ❑ For accelerated development in language development, **non-restrictive open access data exchange** is the way forward!

THANK YOU



<https://ghananlp.github.io/>



natural.language.processing.gh@gmail.com