**Data Sourcing & Preprocessing Documentation**

**What Dataset used:**The dataset used in the **Ehn-bible dataset**, sourced from a project aimed at helping machines understand Nigerian Pidgin English. This dataset is a parallel corpus containing over **10,000 sentences** translated from standard English to Nigerian Pidgin. It was created to support tasks in **machine translation**, **natural language processing (NLP)**, and **language modeling**, especially for under-resourced African languages.

### **Why this dataset was chosen:**

**Focus on Nigerian Pidgin**: Nigerian Pidgin is widely spoken but underrepresented in AI and NLP models. This dataset helps bridge the gap by providing high-quality translations between English and Pidgin.  
**Enables inclusive AI development**: Using this dataset supports the creation of inclusive and locally relevant AI tools, such as chatbots or voice assistants, that can understand and communicate in Nigerian Pidgin.  
**Open-source and curated**: The dataset was carefully curated and is openly available via Hugging Face, which makes it accessible and reliable for training and evaluation purposes .

**Preprocessing:** **Step-by-Step**

Upload File to Google Colab

Select and upload your file when prompted.

Install & Import Required Libraries

Load the Excel File and Show first few rows

Check for null values and Drop rows with nulls

Clean & Normalize the Text: Normalize to lowercase,Remove numbers,Remove punctuation,Remove extra spaces.

Show Basic Text Statistics: Word count per sentence,Sentence count (assuming 1 sentence per row here),Show summary,Show most common words.

**My challenges and observation:**

Challenge was more on how to download dataset from the hugging face and File Upload Issues on google colab. Observation saw a lot of dataset that was clean already and required minimal pre-processing.