3. Dataset

We begin by investigating the dataset that will be used to train and evaluate NLP pipeline. The dataset used for **Neural Machine Translation on Nepali Texts** is available at corpus.zip. However, that will take a long time to train a neural network on and also for preprocessing. Therefore, We used an identical but smaller dataset we found on GitHub: Small_Dataset.zip that contains a small vocabulary so that we will be able to train our model in a reasonable time with this dataset and will be helpful for model experimentation. Then we will use the original dataset after finding the right model for training and testing.

The data is located in `data/small_vocab_en` and `data/small_vocab_ne`. The `small_vocab_en` file contains English sentences with their nepali translations in the `small_vocab_ne` file. Each line in `small_vocab_en` contains an English sentence with the respective nepali translation in each line of `small_vocab_ne`.

```
small_vocab_en Line 1:  It happened after the death of Saul, when David was returned from the slaughter of the Amalekites, and
David had stayed two days in Ziklag;
small_vocab_ne Line 1:  दाऊदले अमालेकीहरूलाई हराएर पछि सिकलग गए। यो शाऊलको मृत्यु भएको केही दिन पछिको कुरा हो। दाऊद त्यहाँ दुइ दिन बसे।
small_vocab_en Line 2:  it happened on the third day, that behold, a man came out of the camp from Saul, with his clothes torn,
and earth on his head: and so it was, when he came to David, that he fell to the earth, and showed respect.
small_vocab_ne Line 2:  तब तेस्रो दिनमा एउटा जवान सैनिक सिकलगमा आयो। त्यो मानिस शाऊलको छाउनीबाट आएको थियो। त्यसका लुगाहरू च्यातिएको र शिर
मा मैला लागेको थियो। त्यसले दाऊदको अघि धोट्टो परेर उनलाई सम्मान गर्न दण्डवत् गर्यो।
small_vocab_en Line 3:  David said to him, "Where do you come from?" He said to him, "I have escaped out of the camp of Israe
l."
small_vocab_ne Line 3:  दाऊदले त्यसलाई सोधे, “तिमी कहाँबाट आयौ?” त्यस मानिसले जवाफ दियो, “म इस्राएली पालबाट आउँदैछु।”
```

Figure 2. First two line of original data

3.1 Vocabulary

The complexity of the problem is determined by the complexity of the vocabulary. A more complex vocabulary is a more complex problem. Therefore, we looked at the complexity of the dataset.

```
2412257 English words.
119048 unique English words.
10 Most common words in the English dataset:
"the" "of" "to" "and" "in" "a" "is" "be" "for" "will"

2160294 Nepali words.
200823 unique Nepali words.
10 Most common words in the Nepali dataset:
"।" "र" "गर्न" "छ" "पनि" "गर्नुहोस्" "अनि" "यो" "लागि" ","
```

Figure 3. Vocabulary