POS Taggers for Indian Languages

In [1]:
```
# Install required libraries (run once)
!pip install indic-nlp-library
import nltk
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
```

```
Unable to create process using 'C:\Users\Yash Dhumal\anaconda3\python.exe "C:\Users
\Yash Dhumal\anaconda3\Scripts\pip-script.py" install indic-nlp-library'
[nltk_data] Downloading package punkt to C:\Users\Yash
[nltk_data]     Dhumal\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\Yash Dhumal\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

Out[1]: True

In [2]:
```
# Importing libraries
from indicnlp.tokenize import indic_tokenize
import nltk
from nltk.tag import UnigramTagger
from nltk.tag import BigramTagger
from nltk.corpus import indian
from nltk.corpus.reader import ConllCorpusReader
```

In [3]:
```
# Sample Hindi sentence
sentence = "मैं स्कूल जा रहा हूँ।"
tokens = list(indic_tokenize.trivial_tokenize(sentence, lang='hi'))
print("Tokenized sentence:", tokens)
```

```
Tokenized sentence: ['मैं', 'स्कूल', 'जा', 'रहा', 'हूँ', '।']
```

In [4]:
```
# Sample tagged corpus (for demo purposes - ideally use a real Hindi POS tagged dat
# Here we simulate a small training corpus
train_data = [
    [('मैं', 'PRP'), ('स्कूल', 'NN'), ('जा', 'VM'), ('रहा', 'VAUX'), ('हूँ', 'VAUX')],
    [('वह', 'PRP'), ('घर', 'NN'), ('गया', 'VM')],
    [('हम', 'PRP'), ('खेल', 'NN'), ('रहे', 'VAUX'), ('थे', 'VAUX')],
]
```

In [5]:
```
# Train a UnigramTagger with backoff to default tagger
default_tagger = nltk.DefaultTagger('NN')
unigram_tagger = UnigramTagger(train_data, backoff=default_tagger)
bigram_tagger = BigramTagger(train_data, backoff=unigram_tagger)
```

In [6]:
```
# Tag a new sentence
tagged = bigram_tagger.tag(tokens)
print("\nPOS Tagged Sentence:")
for word, tag in tagged:
    print(f"{word} --> {tag}")
```

```
POS Tagged Sentence:
मैं --> PRP
स्कूल --> NN
जा --> VM
रहा --> VAUX
हूँ --> VAUX
। --> NN
```

In [ ]: