Descriptive Analysis of Hindi Data

Linguistic Data 3: Data Modeling in ILs

Objective

The objective of this assignment is to perform a **descriptive linguistic and statistical analysis** of Hindi dependency-annotated data from the **Hindi–Urdu Treebank (HUTB)**. Students should submit a report for evaluation containing the following via the moodle folder setup for this purpose:

1. Answers to each question posed (PDF format)
2. github link containing the code and README file

Tasks

The marks for each question is provided in parentheses at the end of the question statement.

1. Basic Corpus Statistics

Compute and report the following:

(a) Total number of sentences (1)

(b) Total number of word tokens (excluding punctuation) (1)

(c) Total number of word types (excluding punctuation) (1)

(d) Average sentence length (in tokens) (1.5)

(e) Minimum and maximum sentence length (1.5)

2. Word Order Patterns

Using dependency relations:

(a) Identify: Subject (`k1`), Object (`k2`), and Main verb (`main`). (3)

(b) Determine the frequency of word order patterns for all the sentences (e.g., SOV, SVO, OSV). (4)

(c) Comment on the dominance of SOV order and the presence of non-canonical orders. (4)

3. Case Marker and Vibhakti Analysis

Case information is encoded in the morphological feature column.

(a) Provide a frequency distribution of case markers in the corpus (3)

(b) The number of unmarked nouns (i.e. nouns without any case marker) (2)

4. Intervening Distance Analysis

(a) For each sentence, compute the average number of intervening words (excluding punctuation) between occurrences of **case markers**. (5)

(b) For each sentence, compute the average number of intervening words (excluding punctuation) between successive occurrences of the **same case marker** and **different case markers**. (5)

(c) Briefly discuss whether certain case markers tend to cluster closely or appear far apart. (3)

5. POS Tag Distribution

Using the fine-grained POS tag column:

(a) Compute the frequency of major POS categories: `NN`, `NNP`, `VM`, `JJ`, `PRP` etc. (3)

(b) Calculate the proportion of verbs relative to nouns. (2)

Evaluation Criteria

- Accuracy of analysis
- Clarity of explanation
- Linguistic insight