Assignment - 1: Descriptive Analysis of Hindi Dependency Data - Report

Linguistic Data 3: Data Modeling in ILs

Vivek Hruday Kavuri
2022114012
Code link: `https://github.com/vivekhruday05/LD3-S-26-Assignments/tree/main/1`

January 25, 2026

Abstract

This report presents a descriptive linguistic and statistical analysis of the Hindi-Urdu Treebank (HUTB). Using a custom Python implementation, we analyzed 20,871 sentences from the InterChunk data to determine basic corpus statistics, word order typology, case marking patterns, intervening distance between markers, and Part-Of-Speech (POS) distributions. A notable finding is the statistical dominance of SVO order (58.17%) over the canonical SOV order (38.42%) in this specific dataset.

1 1. Basic Corpus Statistics

1.1 Implementation Details

The analysis script iterates through the CoNLL-formatted data files. To ensure accuracy and avoid double-counting, the script was restricted to load only the `InterChunk/CoNLL/wx` directory.

- **Sentence Count:** Incremented upon encountering a blank line that signifies the end of a sentence block.
- **Token Count:** Calculated by summing the number of non-punctuation tokens. Tokens with the POS tag `SYM` (Symbol/Punctuation) were explicitly excluded.
- **Word Types:** A Python `set` was used to store unique word forms (excluding punctuation) to determine the vocabulary size.

1.2 Results

Table 1: Basic Corpus Statistics

Metric	Value
Total Number of Sentences	20,871
Total Word Tokens (excluding punctuation)	213,370
Total Word Types (Vocabulary Size)	16,980
Average Sentence Length	10.22 tokens
Minimum Sentence Length	1 token
Maximum Sentence Length	57 tokens

2 2. Word Order Patterns

2.1 Algorithm & Implementation

Hindi is linguistically classified as an SOV language, but it exhibits relatively free word order. To accurately capture the effective word order, we implemented a **Dependency-Aware Surface Order Extraction** algorithm:

1. **Index Mapping:** A hash map is built for each sentence to map every Token ID to its linear index (position).
2. **Predicate Identification (V):**
   - The algorithm identifies the main verb head using `deprel='main'`.
   - **Verb Complex Handling:** The algorithm searches for all children of the main verb that are auxiliaries (POS tag starting with 'V').
   - **Surface Position:** The **rightmost** index (maximum linear position) among the main verb and its auxiliaries is selected as the representative position of the Verb (V_{pos}).
3. **Argument Identification (S & O):**
   - **Subject (S):** Identified by dependency label `k1`.
   - **Object (O):** Identified by dependency label `k2`.
   - **Constraint:** To handle complex sentences, the algorithm strictly checks `parent_id`. Only Subjects and Objects that are **direct children** of the identified Main Verb are counted.
4. **Pattern Determination:** The indices of S, O, and V_{pos} are sorted to produce the final string pattern.

2.2 Results

The analysis of sentences containing a Subject, Object, and Main Verb reveals the following distribution:

Table 2: Frequency of Word Order Patterns

Pattern	Count	Percentage
SVO	3,907	58.17%
SOV	2,580	38.42%
OSV	221	3.29%
Other (OVS, VSO)	8	0.12%

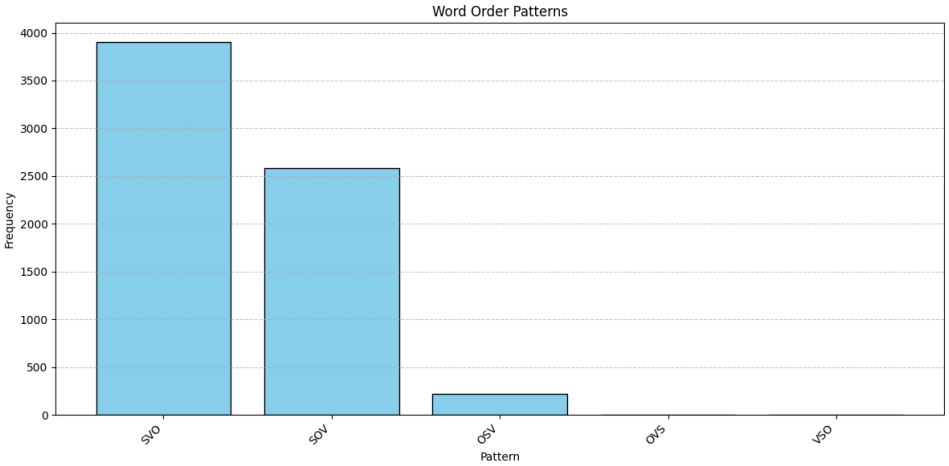

Figure 1: Distribution of Word Order Patterns showing SVO dominance

2.3 Discussion

While Hindi is canonically SOV, the data reveals a dominance of **SVO (58.17%)**. Because our algorithm explicitly handles the verb complex, this suggests that objects (especially clausal complements or heavy NPs labeled as `k2`) are frequently extraposed to the post-verbal position in this specific corpus.

3 3. Case Marker and Vibhakti Analysis

3.1 Algorithm & Implementation

Case markers (Vibhaktis) were extracted from the morphological feature column.

- **Extraction:** We used Regex (`vib-([^|]+)`) to extract the ‘vib‘ attribute.
- **Unmarked Noun Detection:** A noun (POS `NN`) was classified as "Unmarked" if the vibhakti was empty/null (’0’) or the morphological case state was explicitly `case-d` (Direct Case).

3.2 Results

The corpus shows a high frequency of Genitive (`kA`) and Locative (`meM`) markers.

Table 3: Top 8 Case Markers (Vibhaktis)

Marker	Count	Percentage
0_kA (Genitive)	21,723	17.63%
0_meM (Locative)	11,417	9.26%
0_ko (Acc/Dat)	8,428	6.84%
yA	8,049	6.53%
0_ne (Ergative)	6,876	5.58%
0_se (Instr/Abl)	5,803	4.71%
hE	4,924	4.00%
0_para (Locative)	4,272	3.47%

Unmarked Nouns:

- **Count:** 48,617
- **Percentage:** 42.27% of all nouns are unmarked (Direct Case).

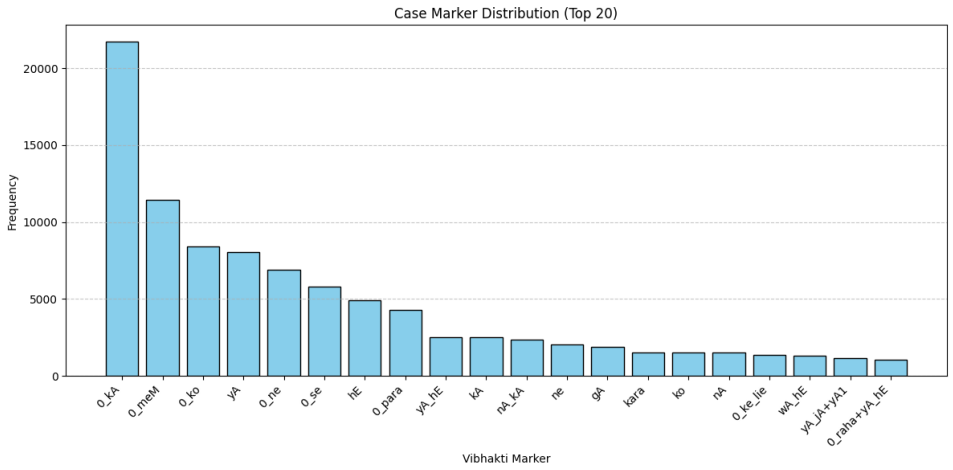

Figure 2: Frequency Distribution of Case Markers

4 4. Intervening Distance Analysis

4.1 Algorithm & Implementation

We calculated the linear distance (number of intervening tokens, excluding punctuation) between successive case markers in a sentence.

1. **Distance Calculation:** For sorted markers (M_i, M_{i+1}), distance $D = Index(M_{i+1}) - Index(M_i) - 1$.
2. **Categorization:** Pairs were split into "Same Marker" (e.g., ne...ne) and "Different Marker" (e.g., ne...ko).

4.2 Results

- **Average Intervening Distance:** 0.74 words (SD: 1.02)

- **Same Marker Distance:** 0.84 words
- **Different Marker Distance:** 0.74 words

4.3 Discussion

The low average distance (0.74) indicates a high density of case-marked tokens. Consistent with linguistic structure, **Different Markers (0.74)** appear closer together than **Same Markers (0.84)**, reflecting the tendency of Hindi sentences to cluster distinct argument roles (e.g., Agent-Object) rather than repeating the same role.

5 5. POS Tag Distribution

5.1 Algorithm & Implementation

POS tags were counted using strict filtering to avoid double-counting (e.g., excluding `NNP` from `NN`).

- **Categories:** NN (Common Noun), NNP (Proper Noun), VM (Main Verb), etc.
- **Ratio Calculation:** $\frac{Total\_VM}{Total\_NN+Total\_NNP}$.

5.2 Results

Table 4: Major POS Category Distribution

Category	Count	Percentage
NN (Common Noun)	85,547	40.09%
VM (Main Verb)	46,618	21.85%
NNP (Proper Noun)	29,440	13.80%
PRP (Pronoun)	17,596	8.25%
CC (Conjunction)	16,963	7.95%
JJ (Adjective)	9,790	4.59%
PSP (Postposition)	385	0.18%

Note on PSP: The low count of explicit Postposition (PSP) tokens (0.18%) contrasts with the high frequency of case markers in Section 3. This indicates that in this dataset, most postpositions are encoded morphologically (in the `vib` column) or attached to the noun, rather than annotated as separate syntactic tokens.

Verb-to-Noun Ratio:

- Total Nouns (NN+NNP): 115,020
- Total Verbs (VM): 46,618
- **Ratio:** 0.41

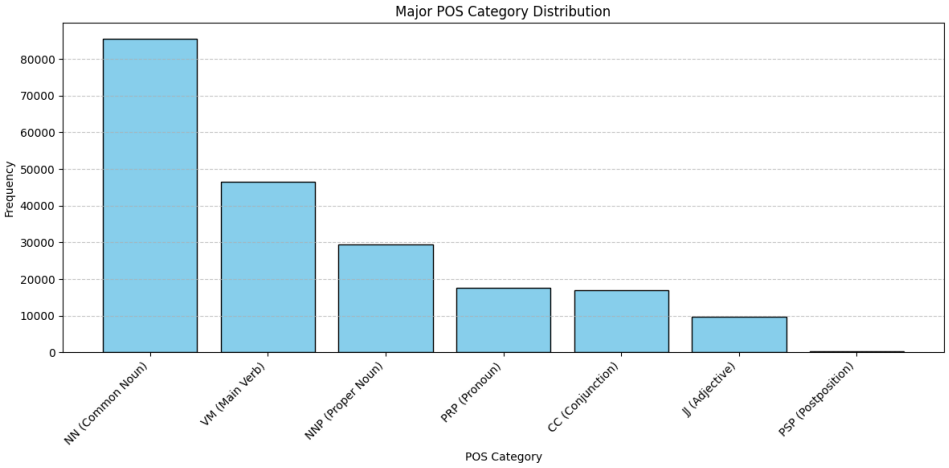

Figure 3: Distribution of Major POS Categories