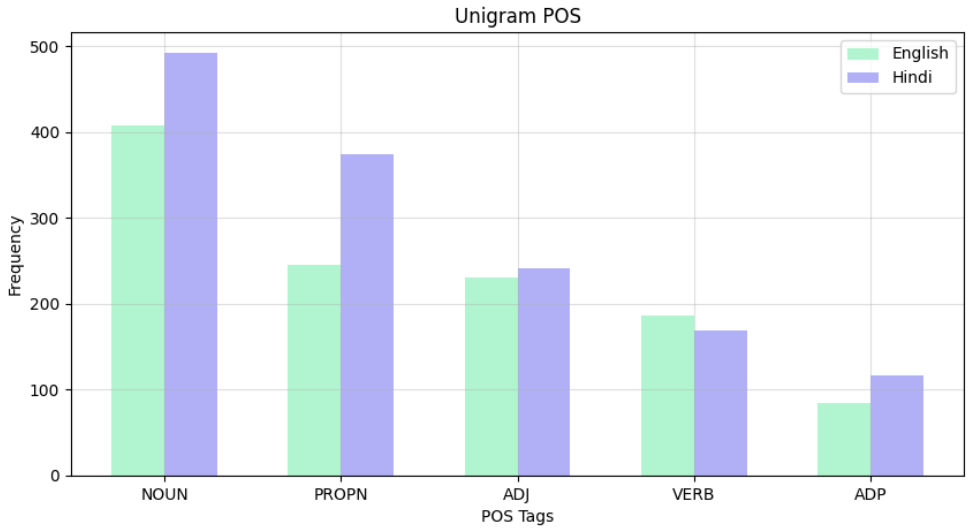
Assignment 3

Deadline	@Mar 2, 2021
Materials	Assignment 3.pdf
Status	Working

Report

Shivansh S, 219114003

Unigram

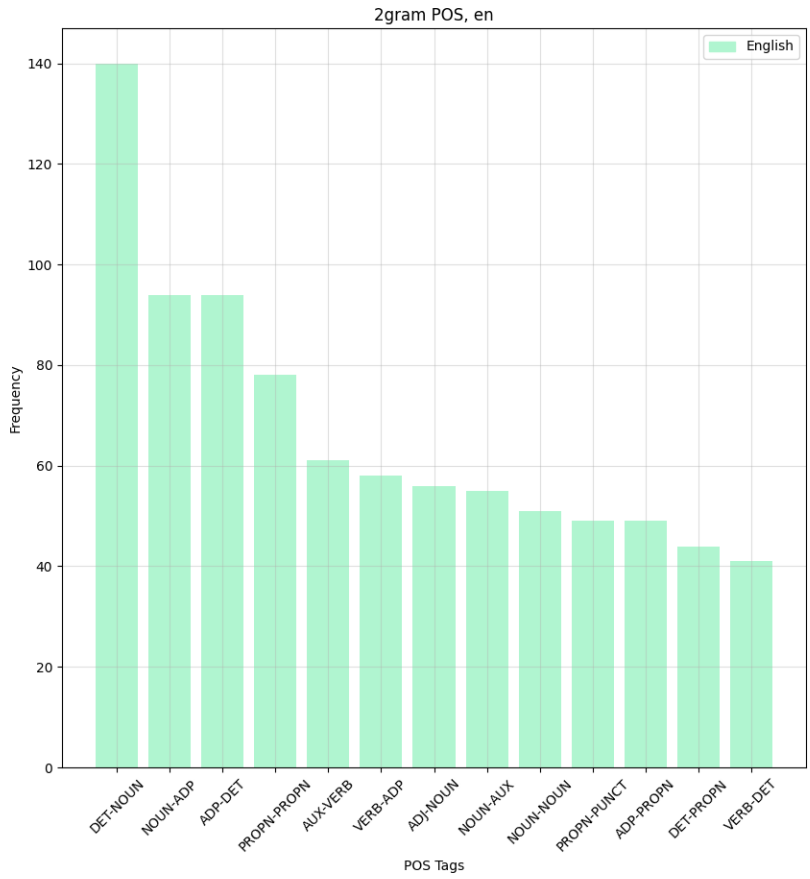


Explanation: The following graph compares the occurrences of particular POS tags in each languages. Since it is parallel corpus, hence it is to be expected that the frequency of occurrences of most tags are almost equal. Here I have only taken the 5 most significant UPOS tags

Observation: As can be seen, the frequency of occurrence of Adjectives and Verbs are almost similar, which is to be expected. What is significant to note is

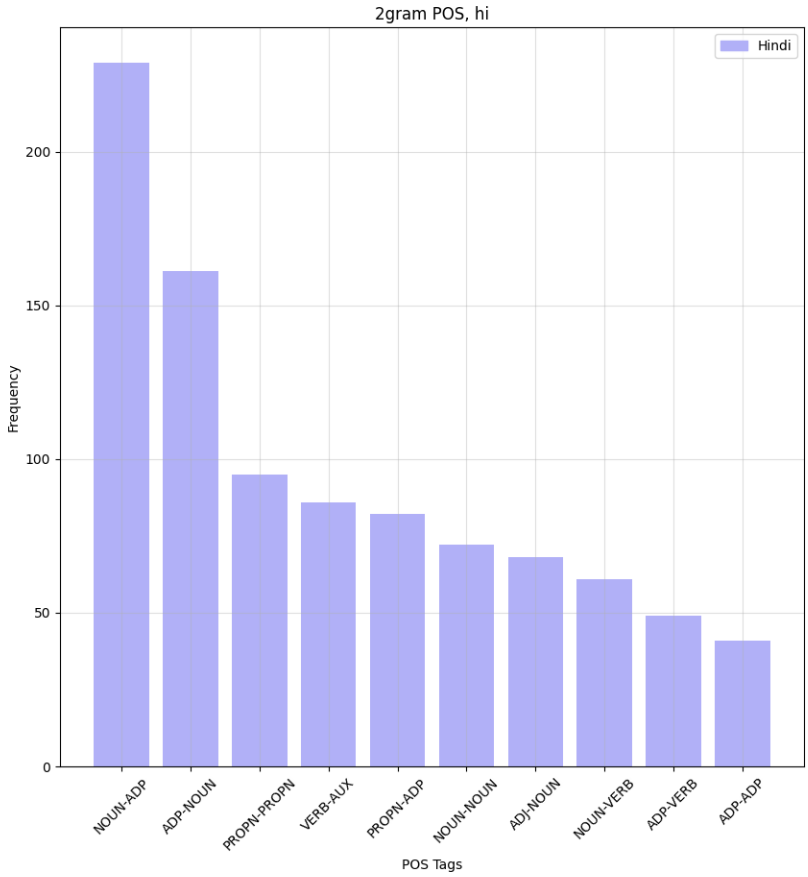
that occurrences of Nouns and Proper Nouns are hugely skewed towards Hindi, whereas occurrences of Adpositions slightly favours Hindi as well.

Bigrams



Explanation: The following graph contains frequency of occurrences of bigrams of POS tags in English. Here, a threshold is set to 40 for aesthetic purposes.

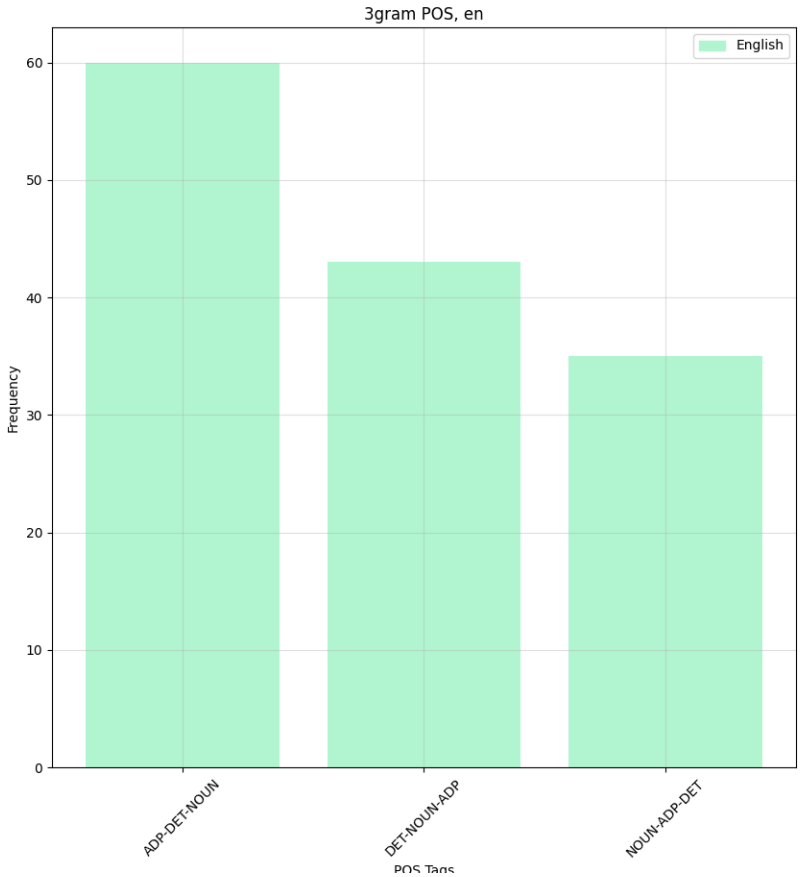
Observation: As expected, we can see that DET is almost always followed by NOUN or PROPN. Another fairly staple observation is that AUX is followed by the VERB, and ADJ by NOUN. An interesting observation is that ADP precedes DET, that is, its considering cases like (..into the Jar). Other observations reflect the structure of the sentence, like NOUN-ADP when you have structure like NOUN-ADP-NOUN etc.



Explanation: The following graph contains frequency of occurrences of bigrams of POS tags in Hindi. Here, a threshold is set to 40 for aesthetic purposes.

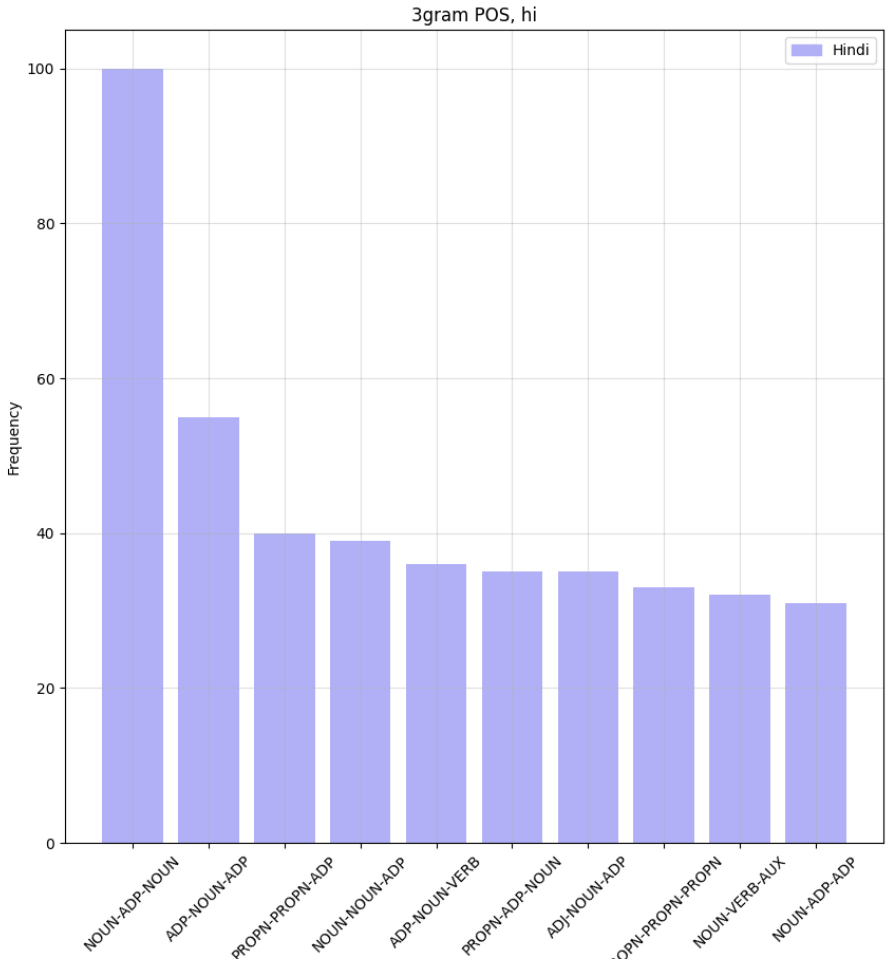
Observation: Here we can see that NOUN-ADP and ADP-NOUN occur with huge frequency, which goes in hand with how postpositions occur in Hindi (explained with graph later). Fairly trivial observations are of VERB-AUX instead of AUX-VERB in English, ADJ-NOUN which is the same as that of English.

Trigrams



Explanation: The following graph contains frequency of occurrences of trigrams of POS tags in English. Here, a threshold is set to 30 for aesthetic purposes.

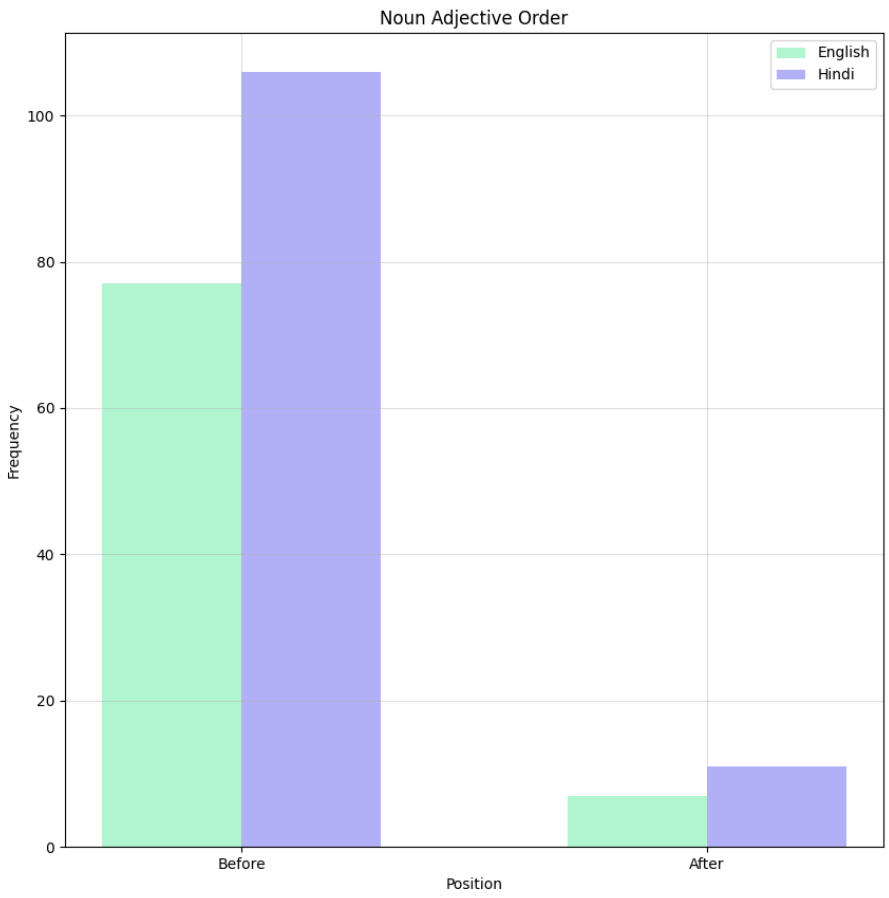
Observation: It is surprising to see that in a fixed order language such as English, we have only 3 patterns which are visible. The ones which are are fairly trivial, and refer to sentences of the structure DET-NOUN-ADP-DET-NOUN.



Explanation: The following graph contains frequency of occurrences of trigrams of POS tags in Hindi. Here, a threshold is set to 30 for aesthetic purposes.

Observation: Again, it is surprising that in a Free Order language, we have so many trigrams who satisfy the frequency threshold. The most frequent one NOUN-ADP-NOUN is fairly trivial, but the interesting one is ADP-NOUN-ADP which signify a difference in structure in Hindi as compared to English. It is interesting to note that almost all patterns contain ADP, hence signifying that there are multiple possible patterns possible when it comes to ADP in Hindi, and the frequency of occurrences. NOUN-NOUN-ADP makes a lot of sense as in the sentence like "ball ko kettle ke andar".

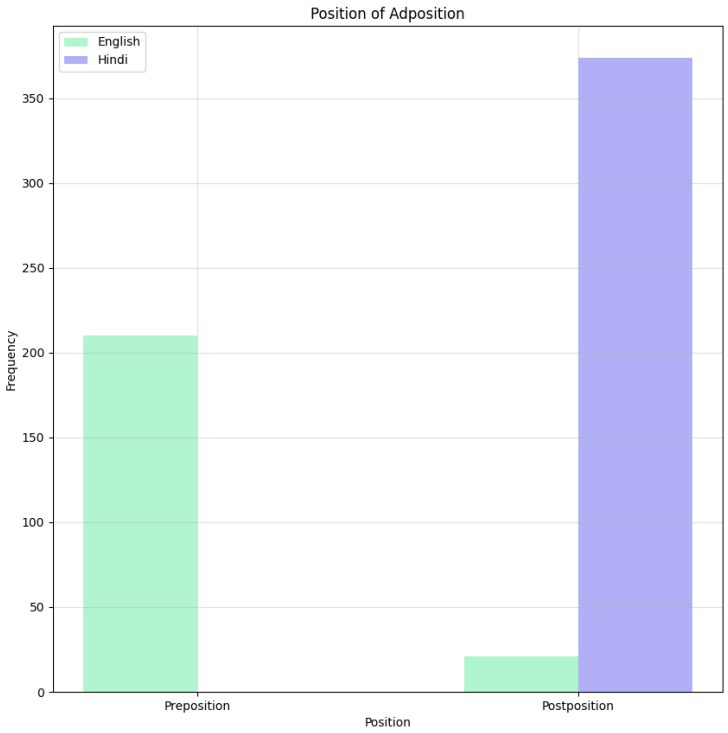
Noun-Adjective



Explanation: The graph contains frequency of Noun-Adjective order in both languages and compares them. Before signifies Adjective occurs before the Noun, and After is the opposite.

Observation: As was seen via the Bigram graph, the Adjective occurs before the Noun in both Hindi and English.

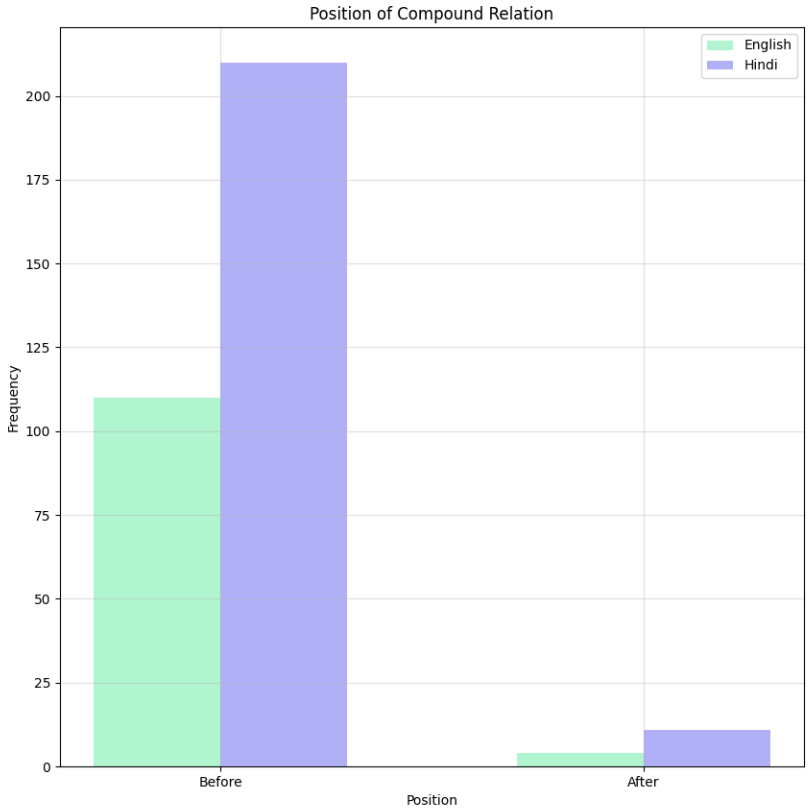
Adpositions



Explanation: It compares the frequency and position of adposition with respect to the Head of the word.

Observation: As expected, majority of the English had preposition, and Hindi had postpositions. It is however, important to note that Hindi had almost 150 more occurrences of ADP compared to English, which is in line with the Unigram Graph.

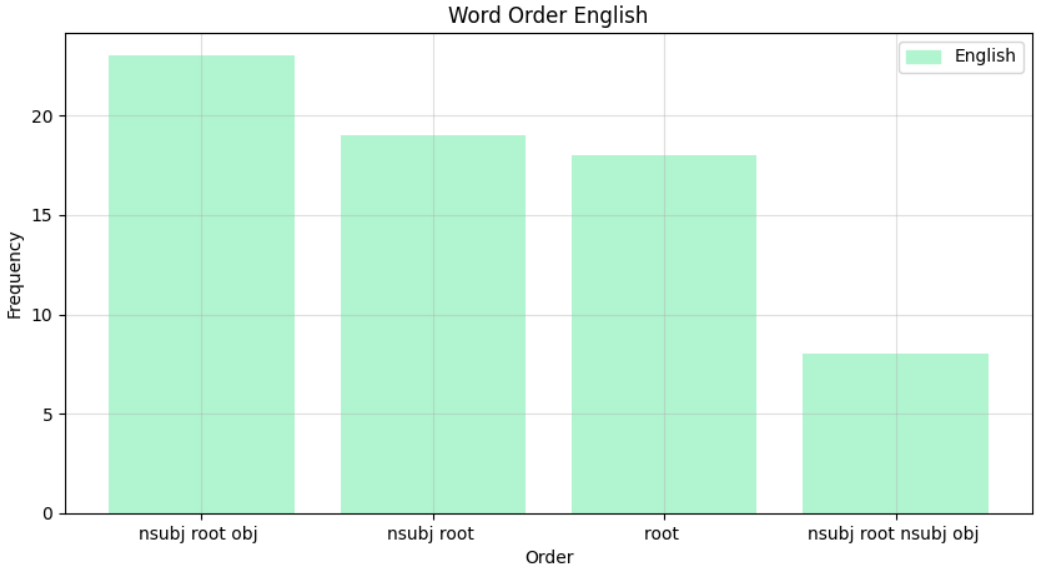
Compound Relation

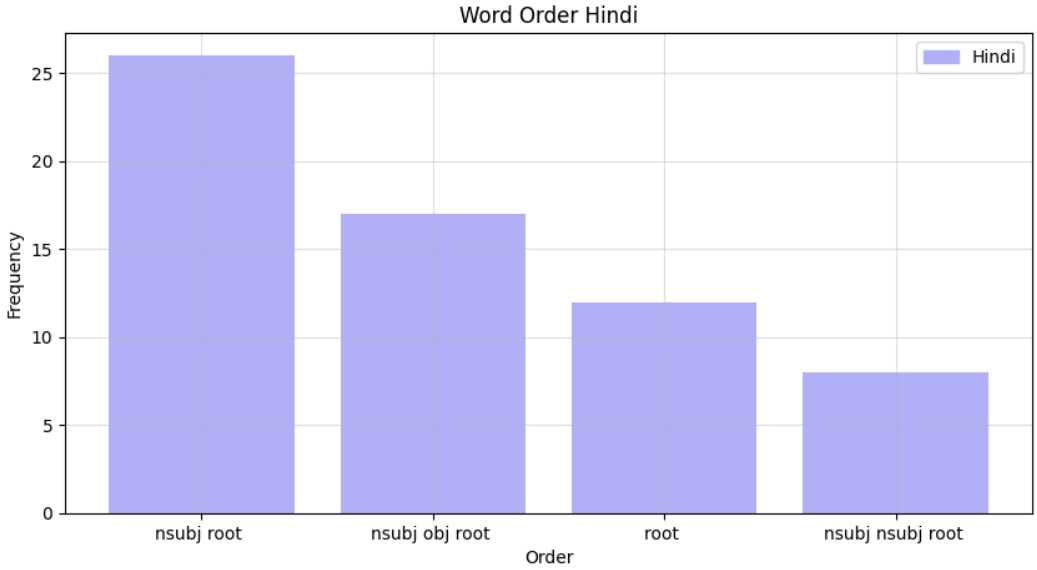


Explanation: Here the graph compares how compound relations occur with respect to the head in English and Hindi.

Observation: It is seen that the compound word occurs before the head word in both the languages. Though it is interesting to see that the frequency of occurrences in Hindi is still way greater than that of English, which might be due to worse model, or case marking considering it as compound instead of separate.

Word Order





Explanation: The graphs above compare the various word orders seen in both languages

Observation: As can be seen, English has SVO word order, and Hindi has SOV. This is fairly routine. Its also seen that Hindi follows the Greenberg's Universals by having postpositions as well. The varying patterns represent sentences with verbs with different transitivity.