Urvashi Senha, 201225020
G Drushti Apoorva, 201225011

Module-5 : Chunking

Data used:

We used the gold POS tagged data wherein there were files for Hindi as well as English sentences. Each line had one sentence POS tagged words.
The tag and the word were separated by an '_'.

The data consisted of twenty sentences for training in English as well as Hindi. Similarly testing data consisted of ten sentences each for Hindi and English.

Procedure for Chunking:

The training sentences were manually analysed and intuitive chunks were made. According to this analysis, the rules were devised.

Tagset used:

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol

25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb
37.	PSP	Post-position
38.	PSPP	Post-positional Phrase
39.	JJP	Adjectival Phrase
40.	RPP	Particle Phrase
41.	RBP	Adverbial Phrase

Observations:

- We have avoided redundancy of rules.
- We have taken care that the rules are as condensed as possible and hence never is there a rule where one tag goes to just one another.
- We have taken into consideration sentences that contain sub-sentences.
- It has been observed that English makes use of Preposition while Hindi uses Postpositions.
- To properly tag postpositional phrases, we have introduced the PSPP chunk.
- As Hindi is a language that allows free-word order, there can be many possibilities for each chunk.