AnnCorra : Annotating Corpora
Guidelines For POS And Chunk Annotation For Indian Languages

Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, Rajeev Sangal
Language Technologies Research Centre
IIIT, Hyderabad

Date : 15-12-2006

OUTLINE

1. Introduction

The significance of large annotated corpora in the present day NLP is widely known. Annotated corpora serve as an important tool for investigators of natural language processing, speech recognition and other related areas. It proves to be a basic building block for constructing statistical models for automatic processing of natural languages.

Many such corpora are available for languages across the world and have proved to be a useful step towards natural language processing. Coming to the scenario for Indian languages, not much work has been carried out on the front of automatic processing of Hindi or any other Indian language. The main bottleneck being unavailability of an annotated corpora, large enough to experiment statistical algorithms.

Annotation of corpora (AnnCorra) can be done at various levels viz, part of speech, phrase/clause level, dependency level, etc. Part of speech tagging forms the basic step towards building an annotated corpus. Chunking can form the next level of tagging.

The task of annotating corpora of several Indian languages has been taken up in the Indian Language Machine Translation (ILMT) project. ILMT is a project in which a number of institutes have come together to form a consortium and work towards developing MT systems for various Indian language pairs.

A primary requirement of such an effort is defining standards for various sub tasks. Thus, standardization of annotation schemes for various annotation tasks becomes a crucial step in this direction.

The issues related to defining standards for POS/Chunk tagging schemes were discussed by scholars from various Indian institutes by way of holding meetings etc. and some standards have been arrived at.

2. Objective

The purpose of the meetings was to arrive at standard tagging scheme for POS tagging and chunking for annotating Indian languages (AnnCorra) and come up with the tags which are exhaustive for the task of annotation for a larger group of languages, specially, Indian languages. The present document gives a detailed description of the tags which have been defined for the tagging schemes and elaborates the motivations behind the selection of these tags. The document also discusses various issues that were addressed while preparing the tag sets and how they have been resolved.

3. Some Assumptions

3.1 During the workshop it was decided to base the discussion and decisions about various tags on the following basic assumptions which everybody agreed on :

i) The tags should be common for all Indian languages
ii) It should be comprehensive/ complete
iii) It should be simple. Maintaining simplicity is important for the following two reasons :
(a) Ease of Learning
(b) Consistency in annotation

3.2 Another important point which was discussed and agreed upon was that POS tagging is NOT a replacement for morph analyser. A 'word' in a text carries the following linguistic knowledge

a) grammatical category and
b) grammatical features such as gender, number, person etc. The POS tag should be based on the 'category' of the word and the features can be acquired from the morph analyser.

4. Issues in Tag Set Design

This section deals with some of the issues related to any POS tagger and the policy that we have adopted to deal with each of these issues for our purpose.

The first step towards developing POS annotated corpus is to come up with an appropriate tags. The major issues that need to be resolved at this stage are :

1. Fineness vs Coarseness in linguistic analysis
2. Syntactic Function vs lexical category
3. New tags vs tags close to existing English tags

4.1 Fineness vs Coarseness

An issue which always comes up while deciding tags for the annotation task is whether the tags should capture 'fine grained' linguistic knowledge or keep it 'coarse'. In other words, a decision has to be taken whether or not the tags will account for finer distinctions of the parts of speech features. For example, it has to be decided if plurality, gender and other such information will be marked distinctly or only the lexical category of a given word should be marked.

It was decided to come up with a set of tags which avoids 'finer' distinctions. The motivation behind this is to have less number of tags since less number of tags lead to efficient machine learning. Further, accuracy of manual tagging is higher when the number of tags is less.

However, an issue of general concern is that in an effort to reduce the number of tags we should not miss out on crucial information related to grammatical and other relevant linguistic knowledge which is encoded in a word, particularly in agglutinating languages, eg, Tamil, Telugu and many other Indian languages. If tags are too coarse, some crucial information for further processing might be missed out. As mentioned above, primarily the required knowledge for a given lexical item is its grammatical category, the features specifying its grammatical information and any other information suffixed into it. For example,

Telugu word ' *rAmudA* (Is it Ram ?)' contains the following information <category (noun)+grammatical features(masculine, singular) + question>. The word by itself is a bundle of linguistic information. Morph analyser provides all the knowledge that is contained in a word. It was decided that any linguistic knowledge that can be acquired from any other source (such as morph analyser) need not be incorporated in the POS. As mentioned above, POS tagger is not a replacement for morph analyser. In fact, features from morph analyser can be used for enhancing the performance of a POS tagger. The additional knowledge of a POS given by a POS tagger can be used to disambiguate the multiple answers provided by a morph analyser.

On the other hand, we agree that too coarse an analysis is not of much use. Essentially, we need to strike a balance between fineness and coarseness. The analysis should not be so fine as to hamper machine learning and also should not be so coarse as to miss out important information. It is also felt that fine distinctions are not relevant for many of the applications(like sentence level parsing, dependency marking, etc.) for which the tagger may be used in future.

However, it is well understood that plurality and other such information is crucial if the POS tagged corpora is used for any application which needs the agreement information. In case such information is needed at a later stage, the same tag set can be extended to encompass information such as plurality etc as well. This can be done by providing certain heuristics or linguistic rules.

Thus, to begin with, it has been decided to adopt a coarse part of speech analysis. At the same time, wherever it is found essential, finer analysis is incorporated. Also, there is a basic understanding that wherever/whenever essential, the tags containing finer linguistic knowledge can be incorporated. An example of where finer analysis becomes crucial has been given below. Take the Hindi sentence (h1) below :

h1. *AsamAna*_NN *meM*_PSP *uDane*_VM *vAlA*_PSP *ghoDA*_NN
'sky' 'in' 'flying' 'horse
*nIce*_NST *utara*_VM *AyA*_VAUX.
'down' “descend” “came”

In (h1) above *uDane* is a noun derived from a verb. The word *AsamAna* is an argument of *uDane* and not of *'nIce utara AyA* – another verb in the sentence. It is crucial to retain the information that *uDane,* though functioning as a noun now, is derived from a verb and can take its own arguments. In order to preserve such crucial information a finer analysis is essential. Therefore, a distinct tag needs to be introduced for such expressions. In the current tagging scheme *uDane* will be annotated as a 'main verb (VM)' at the POS level. However, the information that it is functioning like a noun will be captured at the chunking level by introducing a distinct chunk tag VGNN (discussed in details under Section III on Chunking).

4.2 Syntactic Function vs Lexical Category

A word belonging to a particular lexical category may function differently in a given context. For example, the lexical category of *harijana* in Hindi is a noun . However, functionally, *harijana* is used as an adjective in (h2) below,

h2. *eka dina pAzca baje khabara AyI ki koI* ***harijana***
'one' 'day' 'five' 'o'clock' 'news' 'came' 'that' 'some' 'harijana'
bAlaka *unase milanA cAhatA hE*

'young boy'   'him'   'to meet'   'wants'   'is'
“One day, a message came at five o'clock that some 'harijana' boy wanted to meet him”.

Such cases require a decision on whether to tag a word according to its lexical category or by its syntactic category. Since the word in a context has syntactic relevance, it appears natural to tag it based on its syntactic information. However, such a decision may lead to further complications.

In AnnCorra, the syntactic function of a word is not considered for POS tagging. Since the word is always tagged according to its lexical category there is consistency in tagging. This reduces confusion involved in manual tagging. Also the machine is able to establish a word-tag relation which leads to efficient machine learning.

In short, it was decided that syntactic and semantic/pragmatic functions were not to be the basis of deciding a POS tag.

4.3. New Tags vs Tags from a Standard Tagger

Another point that was considered while deciding the tags was whether to come up with a totally new tag set or take any other standard tagger as a reference and make modifications in it according to the objective of the new tagger. It was felt that the later option is often better because the tag names which are assigned by an existing tagger may be familiar to the users and thus can be easier to adopt for a new language rather than a totally new one. It saves time in getting familiar to the new tags and then work on it.

The Penn tags are most commonly used tags for English. Many tag sets designed subsequently have been a variant of this tag set (eg. Lancaster tag set). So, while deciding the tags for this tagger, the Penn tags have been used as a benchmark. Since the Penn tag set is an established tag set for English, we have used the same tags as the Penn tags for common lexical types. However, new tags have been introduced wherever Penn tags have been found inadequate for Indian language descriptions. For example, for verbs none of the Penn tags have been used. Instead, AnnCorra has only two tags for annotating verbs, VM (main verb) and VAUX (auxiliary verb).

5. POS tags Chosen for the Current Scheme

This section gives the rationale behind each tag that has been chosen in this tag set.

5.1.1 NN Noun

The tag NN for nouns has been adopted from Penn tags as such. The Penn tag set makes a distinction between noun singular (NN) and noun plural (NNS). As mentioned earlier, distinct tags based on grammatical information are avoided in IL tagging scheme. Any information that can be obtained from any other source is not incorporated in the POS tag. Plurality, for example, can be obtained from a morph analyzer. Moreover, as mentioned earlier, if a particular information is considered crucial at the POS tagging level itself, it can be incorporated at a later date with the help of heuristics and linguistic rules. This approach brings the number of tags down, and helps achieve simplicity, consistency, better machine learning with a small corpora etc. Therefore, the current scheme has only one tag (NN) for common nouns without getting into any distinction based on the grammatical information contained in a given noun word

5.1.2 NST Noun denoting spatial and temporal expressions

A tag NST has been included to cover an important phenomenon of Indian languages. Certain expressions such as '*Upara*' (above/up), '*nIce*' (below) '*pahale*' (before), 'Age' (front) etc are content words denoting time and space. These expressions, however, are used in various ways. For example,

5.1.2.1 These words often occur as temporal or spatial arguments of a verb in a given sentence taking the appropriate *vibhakti* (case marker):

h3. *vaha* ***Upara*** *so rahA thA* .
'he' 'upstairs' 'sleep' 'PROG' 'was'
“He was sleepign upstairs”.

h4. *vaha* ***pahale se*** *kamare meM bEThA thA* .
'he' 'beforehand' 'from' ' room' 'in' 'sitting' 'was'
“He was sitting in the room from beforehand”

h5. *tuma* ***bAhara*** *bETho*
'you' 'outside' 'sit'
“You sit outside”.

Apart from functioning like an argument of a verb, these elements also modify another noun taking postposition 'kA'.

h6. *usakA baDZA bhAI* ***Upara ke hisse meM*** *rahatA hE*
'his' 'elder' 'brother' 'upstairs' 'of' 'portion' 'in' 'live' 'PRES'
“His elder brother lives in the upper portion of the house”.

5.1.2.2 Apart from occuring as a nominal expression, they also occur as a part of a postposition along with 'ke'. For example,

h7. *ghaDZe* ***ke Upara*** *thAlI rakhI hE.*
'pot' 'of' 'above' 'plate' 'kept' 'is'
The plate is kept on the pot”.

h8. *tuma ghara ke* **bAhara** *bETho*
'you' 'home' 'of' 'outside' 'sit'
"You sit outside the house".

'Upara' and '*bAhara*' are parts of complex postpositions *'ke Upara'* and '*ke bAhara*' in (h6) and (h7) respectively which can be translated into English prepositions 'on' and 'outside'.

For tagging such words, one possible option is to tag them according to their syntactic function in the given context. For example in 5.2.2 (h7) above, the word '*Upara*' is occurring as part of a postposition or a relation marker. It can, therefore, be marked as a postposition. Similarly, in 5.2.1. (h3) and (h6) above, it is a noun, therefore, mark it as a noun and so on. Alternatively, since these words are more like nouns, as is evident from 5.2.1 above they can be tagged as nouns in all there occurrences. The same would apply to '*bAhAra'* (outside) in examples examples (h4), (h5) and (h8).

However, if we follow any of the above approaches we miss out on the fact that this class of words is slightly different from other nouns. These are nouns which indicate 'location' or 'time'. At the same time, they also function as postpositions in certain contexts. Moreover, such words, if tagged according to their syntactic function, will hamper machine learning. Considering their special status, it was considered whether to introduce a new tag, NST, for such expressions. The following five possibilities were discussed :

a) Tag both (h5) & (h8) as NN
b) Tag both (h5) & (h8) as NST
c) Tag (h5) as NN & (h8) as NST
d) Tag (h5) as NST & (h8) as PSP
e) Tag (h5) as NN & (h8) as PSP

After considering all the above, the decision was taken in favour of (b). The decision was primarily based on the following observations:

(i) *'bAhara'* in both (h5) and (h8) denotes the same expression (place expression 'outside')
(ii) In both (h5) and (h8), *'bAhara'* can take a vibhakti like a noun (**bAhara ko** bETho, ghara **ke bAhara ko** bETho)
(iii) If a single tag is kept for both the usages, the decision making for annotators would also be easier.

Therefore, a new tag **NST** is introduced for such expressions. The tag **NST** will be used for a finite set of such words in any language. For example, Hindi has *Age* (front), *pIche* (behind), *Upara* (above/upstairs), *nIce* (below/down), *bAda* (after), *pahale* (before), *andara* (inside), *bAhara (outside)* etc.

5.2 NNP Proper Nouns

The need for a separate tag for proper nouns and its usability was discussed. Following points were raised against the inclusion of a separate tag for proper nouns :

a) Indian languages, unlike English, do not have any specific marker for proper nouns in orthographic conventions. English proper nouns begin with a capital letter which distinguishes them from common nouns.
b) All the words which occur as proper nouns in Indian languages can also occur as common nouns denoting a lexical meaning. For example,
English : John, Harry, Mary occur only as proper nouns whereas
Hindi : *aTala bihArI, saritA, aravinda* etc are used as 'names' and they also belong to grammatical categories of words with various senses . For example given below is a list of Hindi words with their grammatical class and sense.

aTala	adj	immovable
bihArI	adj	from Bihar
saritA	noun	river
aravinda	noun	lotus

Any of the above words can occur in texts as common lexical items or as proper names. (h9) - (h11) below show their occurrences as proper nouns,

h9. ***atala bihAri bAjapaI*** *bhArata ke pradhAna mantrI the.*
'Atal' 'Bihari' 'Vajpayee' 'India' 'of' 'prime' 'minister' 'was'
"Atal Behari Vajpayee was the Prime Minister of India".

h10. merI mitra **saritA** tAIvAna jA rahI hE.
'my' 'friend' 'Sarita' 'Taiwan' 'go' 'PROG' 'is'
"My friend Sarita is going to Taiwan"

h11. ***aravinda*** *ne mohana ko kitAba dI.*
'Aravind' 'erg' 'Mohan' 'to' 'book' 'gave'
"Aravind gave the book to Mohan".

Therefore, in the Indian languages' context, annotating proper nouns with a separate tag will not be very fruitful from machine learning point of view. In fact, the identification of proper nouns can be better achieved by named entity filters.

Another point that was considered in this context was the effort involved in manual tagging of proper nouns in a given text. It is felt that not much extra effort is required in manual tagging of proper nouns. However, the data annotated with proper nouns can be useful for certain applications. Therefore, there is no harm in marking the information if it does not require much effort.

Finally, it was decided to have a separate tag for proper nouns for manual annotation and ignore it for machine learning algorithms. Following this decision, the tag **NNP** is included in the tag set. This tag is the same as the

Penn tag for proper nouns. However, in this case also AnnCorra has only one tag for both singular and plural proper nouns unlike Penn tags where a distinction is made between proper noun singular and proper noun plural by having two tags NNP and NNPS respectively.

5.3.1 PRP Pronoun

Penn tags make a distinction between personal pronouns and possessive pronouns. This distinction is avoided here. All pronouns are marked as PRP. In Indian languages all pronouns inflect for all cases (accusative, dative, possessive etc.). In case we have a separate tag for possessive pronouns, new tags will have to be designed for all the other cases as well. This will increase the number of tags which is unnecessary. So only one tag is used for all the pronouns. The necessity for keeping a separate tag for pronouns was also discussed, as linguistically, a pronoun is a variable and functionally it is a noun. However, it was decided that the tag for pronouns will be helpful for anaphora resolution tasks and should be retained.

5.3.2 DEM Demonstratives

The tag 'DEM' has been included to mark demonstratives. The necessity of including a tag for demonstratives was felt to cover the distinction between a pronoun and a demonstrative. For example,

h12. ***vaha ladakA*** *merA bhAI* *hE* (demnostrative)
'that' 'boy' 'my' 'brother' 'is'
h13. ***vaha*** *merA bhAI* *hE* (pronoun)
'he' 'my' 'brother' 'is'

Many Indian languages have different words for demonstrative adjectives and pronouns. A better evidence for including a separate tag for demonstratives is from the following Telugu examples,

t1. *A abbAyi nA tammudu*
'that' 'boy' 'my' 'brother'
t2. *atanu nA tammudu*
'he' 'my' 'brother'
(Telugu does not have a copula 'be' in the present tense)

5.4 VM Verb Main

Verbal constructions in languages may be composed of more than one word sequences. Typically, a verb group sequence contains a main verb and one more auxiliaries (V AUX AUX ... ...). In the current tagging scheme the support verbs (such as *dAlanA* in *kara dAlAtA hE*, *uThanA* in *cOMka uThA thA* etc) are also tagged as VAUX. The group can be finite or non-finite. The main verb need not be marked for finiteness. Normally, one of the auxiliaries carries the finiteness feature.

The necessity of marking the finiteness or non-finiteness in a verb was discussed extensively and everybody agreed that it was crucial to mark the distinction. However, languages such as Hindi, which have auxiliaries for marking tense, aspect and modalities pose a problem. The finiteness of a verbal expression is known only when we reach the last auxiliary of a verb group. Main verb of a finite verb group (leaving out the single word verbal expressions of the finite type – eg *vaha dillI gayA*) does not contain finiteness information. For example,

h14. *laDZakA seba* ***khAtA*** ***raHA wA***
'boy' 'apple' 'eating' 'PROG' 'was'
The boy had kept eating.

h15. *seba* ***khAtA*** ***huA*** *laDZakA jA rahA thA*
'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'
The boy eating the apple was going.

The expression *khAtA raHA* in (h29) above is finite and *khAtA huA* in (h3) is non finite. However, the main verb '*khAtA*' is non-finite in both the cases.

So, the issue is - whether to (1a) mark finiteness in *“**khAtA rahA thA (*** had kept eating*)*” at the lexical level on the main verb (khA) or (1b) on the auxiliary containing finiteness (wA) or (2) not mark it at the lexical level at all. All the three possibilities were discussed;
1) Mark the finiteness at the lexical level.

If we mark it at the lexical level, following possibilities are available :

1a) Mark the finiteness on the main verb, even though we know that the lexical item itself is not finite.

In this case, the annotator interprets the finiteness from the context. (The POS tags VF, VNF and VNN were earlier decided based on this approach). The main verb, therefore, is marked as finite consciously with a view that the group contains a 'verb root' and its auxiliaries (as TAM etc) is finite even though the main verb does not carry the finiteness at the lexical level. Although, this approach facilitates annotation of both the main verb and the finiteness (of the group) by a single tag, it allows tagging a lexical item (main verb) with the finiteness feature which it does not actually carry. So, this is not a neat solution.

1b) The second possibility is, mark the finiteness on the last auxiliary of the sequence. Here again the decision has to be taken from the context. This possibility was not considered since this also involves marking the verb finiteness at the lexical level.

2) Don't mark the finiteness at the lexical level. Instead mark it as indicated in (2a) or (2b) below.

2a) Introduce a new layer which groups the verb group and mark the verb group as finite or non-finite. This approach proposes the following :

(i) Annotate the main verb as **VM** (introduce a new tag). Thus,

h14a. *laDZakA seba* ***khAtA*_VM** *raHA thA*
'boy' 'apple' 'eating' 'PROG' 'was'

h15a. *seba* ***khAtA*_VM** *huA laDZakA jA rahA thA*
'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

(ii) Annotate the auxiliaries as **VAUX**,

h14a. *laDZakA seba khAtA*_VM *raHA*_**VAUX** *thA*_**VAUX**
'boy' 'apple' 'eating' 'PROG' 'was'
h15a. *seba khAtA*_VM *huA*_**VAUX** *laDZakA jA rahA thA*
'apple' 'eating' ' PROG' 'boy' ' go' 'PROG' 'was'

(iii) Group the verb group (before chunking) and annotate it as finite or non-finite as the case may be,

h14a. *laDZakA seba* [*khAtA*_VM *raHA*_VAUX *wA*_VAUX**]_VF**
'boy' 'apple' 'eating' 'PROG' 'was'
h15a. *seba* [*khAtA*_VM *huA*_VAUX**]_VNF** *laDZakA jA rahA thA*
'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

This approach is more faithful to the available linguistic information. However, it requires introducing another layer. So, this was not considered useful.

2b) Mark the finiteness at the chunk level,

In this approach, the lexical items are marked as in (2). No new layer is introduced. Instead, the decision is postponed to the chunk level. Since the finiteness is in the group, it is marked at the chunk level. This offers the best solution as it facilitates marking the linguistic information as it is without having to introduce a new layer.

h14a. *laDZakA seba* ((*khAtA*_VM *raHA*_VAUX *wA*_VAUX))**_VGF**
'boy' 'apple' 'eating' 'PROG' 'was'

h15a. *seba* ((*khAtA*_VM *huA*_VAUX))**_VGNF** *laDZakA jA rahA thA*
'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

In this case also the decision is made by looking at the entire group. (2b) was most preferred as it facilitates marking the linguistic information correctly, at the same time no new layer needs to be introduced. Therefore, the current tagging scheme has adopted this approach. Thus, the main verbs in a given verb group will be marked as **VM,** irrespective of whether the total verb group is finite of non finite. Given underneath are some examples of other verb group types :

1) **Non finite verb groups -** Non-finite verb groups can have two functions :

a) Adverbial participial, for example : *khAte-khAte* in the following Hindi sentence,

h16. *mEMne* ***khAte – khAte*** *ghode ko dekhA*
'I erg' 'while eating' 'horse' 'acc' 'saw'
“I saw a horse while eating”.

The main verb in (h16) would be annotated as follows :

h16a. *mEMne* ***khAte – khAte*****_VM** *ghode ko dekhA*

b) Adjectival participial, for example : '*khAte Hue*' in the following Hindi sentence ,

h17. *mEMne ghAsa* ***khAte_VM hue*** *ghoDe ko dekhA* *
'I erg' 'grass' 'eating' 'PROG' 'horse' 'acc' 'saw'
I saw the horse eating grass.

(* (h17) is ambiguous in Hindi. The other sense that it can have is, *I saw the horse while (I was) eating grass*. In such cases, the annotator would disambiguate the sentence depending on the context and mark accordingly.)

2) Gerunds

Functionally, gerunds are nominals. However, even though they function like nouns, they are capable of taking their own arguments,eg. *pInA* in the following Hindi sentence can occur on its own or take an argument (given in parenthesis):

h18. (*sharAba)* ***pInA*****_VM** *sehata ke liye hAnikAraka hE.*
'liquor' 'drinking' 'health' 'for' 'harmful' 'is'
“Drinking (liquor) is bad for health”

h19. *mujhe* ***khAnA*****_VM** *acchA lagatA hai*
'to me' 'eating' 'good' 'appeals'
“I like eating”

h20. ***sunane meM*** *saba kuccha acchA lagatA hE* 'listening' 'in' 'all' 'things' 'good' 'appeal' 'is'

As mentioned above, noun '*sharAba*' in (h18) is an object of the verb '*pInA*' and has no relation to the main verb (*hE*). In order to be able to show the exact verb-argument structure in the sentence, it is essential that the crucial information of a noun derived from a verb is preserved. Therefore, even gerunds have to be marked as verbs. It is proposed that in keeping with the approach adopted for non-finite verbs, mark gerunds also as **VM** at the lexical level. For capturing the information that they are gerunds, such verbs will be marked as **VGNN** (see the section on Chunk tags for details) at the chunk level to capture their gerundial nature. The verbs having 'vAlA' vibhakti will also be marked as VM. For example, '*khonevAlA*' (one who looses).

5.5 VAUX Verb Auxiliary

All auxiliary verbs will be marked as VAUX. This tag has been adopted as such from the Penn tags. (For examples, see h14 – h16 above).

5.6 JJ Adjective

This tag is also taken from Penn tags. Penn tag set also makes a distinction between comparative and superlative adjectives. This has not been considered here. Therefore, in the current scheme for Indian languages, the tag JJ includes the 'tara' (comparative) and the 'tama' (superlative) forms of adjectives as well.
For example, Hindi *adhikatara* (more times), *sarvottama* (best), etc. will also be marked as JJ.

5.7 RB Adverb

For the adverbs also, the tag RB has been borrowed from Penn tags. Similar to the adjectives, Penn tags make a distinction between comparative and superlative adverbs as well. This distinction is not made in this tagger. This is in accordance with our philosophy of coarseness in linguistic analysis.
Another important decision for the use of RB for adverbs in the current scheme is that :-

(a) The tag RB will be used ONLY for 'manner adverbs' . Example,

h21. *vaha* *jaldI jaldI* *khA* *rahA* *thA*
'he' 'hurriedly' 'eat' 'PROG' 'was'

(b) The tag RB will NOT be used for the time and manner expressions unlike English where time and place expressions are also marked as RB. In our scheme, the time and manner expressions such as '*yahAz – vahAz, aba – waba* ' etc will be marked as PRP.

5.8 PSP Postposition

All Indian languages have the phenomenon of postpositions. Postpositions express certain grammatical functions such as case etc. The postposition will be marked as PSP in the current tagging scheme. For example,

h22. m*ohana* *kheta* ***meM*** *khAda* *dAla* *rahA* *thA*
'Mohan' 'field' 'in' 'fertilizer' 'put ' 'PROG' 'was'

meM in the above example is a postposition and will be tagged as PSP.
A postposition will be annotated as PSP ONLY if it is written separately. In case it is conjoined with the preceding word it will not be marked separately. For example, in Hindi pronouns the postpositions are conjoined with the pronoun,

h23. *mE**ne*** *usa**ko*** *bAzAra* ***meM*** *dekhA*
'I' 'him' 'market' 'in' 'saw'

(h23) above has three instances of 'postposition' (in bold) usage. The postpositions '*ne*' and '*ko*' are conjoined with the pronouns *mEM* and *usa* respectively. The third postposition '*meM*' is written separately. In the first two instances, the postposition will not be annotated. Such words will be annotated with the category of the head word. Therefore, the three instances mentioned above will be annotated as shown in (h23a) below :

h23a. *mE**ne***_**PRP** *usa**ko***_**PRP** *bAzAra*_**NN** ***meM***_**PSP** *dekhA*

5.9 RP Particle

Expressions such as *bhI, to, jI, sA, hI, nA,* etc in Hindi would be marked as RP. The *nA* in the above list is different from the negative *nA.* Hindi and some other Indian languages have an ambiguous 'nA' which is used both for negation (NEG) and for reaffirmation (RP). Similarly, the particle *wo* is different from CC *wo*. For example in Bangla and Hindi:

Bangla : (b1) *tumi* ***nA*_RP** *khub* *dushtu*
'you' 'particle' 'very' 'naughty'
“You are very naughty” (comment)

Hindi : (h24) *tuma* ***nA*_RP**, *bahuta dushta ho*
'you' 'particle very naughty
“You are very naughty” (comment)

Bangla : (b2) *cheleta* *dushtu* ***nA*_NEG**
'the boy' 'naughty' 'not'
“The boy is not naughty”
Hindi : (h25) *mEM* ***nA*_NEG** *jA* *sakUMgA*
'I' 'not' 'go' 'will able'
“I will not be able to go”

Bangla : (b3) *binu yYoxi khAya* **to_CC** *Ami khAba*
'Binu' 'if' 'eats' 'then' 'I' 'will eat'
"If Binu eats then I will eat (too)"
Hindi : (h26) *yadi binu khAyegA* **wo_CC** *mEM khAUMgI*
'if' 'Binu' 'eats' 'then' 'I' 'will eat'
"Only if Binu eats, I will eat (too)"
Bangla : (b4) *Ami* ***to*_RP** *jAni nA*
'I' 'particile' 'know' 'not'
"I don't know"
Hindi : (h27) *mujhako* ***to*_RP** *nahIM patA*
'I' 'particile' 'not' 'know'
"I don't know"

5.10 CC Conjuncts(co-ordinating and subordinating)

The tag CC will be used for both, co-ordinating and subordinating conjuncts. The Penn tag set has used IN tag for prepositions and subordinating conjuncts. Their rationale behind this is that subordinating conjuncts and prepositions can be distinguished because subordinating conjuncts are followed by a clause and prepositions by a noun phrase.

But in the current tagger all connectives, other than prepositions, will be marked as CC.

h28. mohana bAzAra jA rahA hE **Ora_CC** ravi skUla jA rahA hE
'Mohan' 'market' 'go' 'PROG' 'is' 'and' 'Ravi' 'school' 'go' 'PROG' 'is'
"Mohan is going to the market and Ravi is going to the school"

h29. mohana ne mujhe batAyA **ki_CC** Aja bAzAra banda hE
'Mohan' 'erg' 'to me' 'told' 'that' 'today' 'market' 'close' 'is'
"Mohan told me that the market is closed today."

5.11 WQ Question Words

The Penn tag set makes a distinction between various uses of 'wh-' words and marks them accordingly (WDT, WRB, WP, WQ etc). The 'wh-' words in English can act as questions, as relative pronouns and as determiners. However, for Indian languages we need not keep this distinction. Therefore, we tag the question words as WQ.

h30. ***kOna*** *AyA hE* ?
'who' 'come' 'has'
"Who has come ?
h31. *tuma kala* ***kyA*** *kara rahe ho* ?
'you' 'tomorrow' what' 'doing' 'are'
What are you doing tomorrow ?

h32. *tuma kala* ***kahAz*** *jA rahe ho* ?
'you' 'tomorrow' 'where' 'going' 'are'
"Where are you going tomorrow ?

h33. ***kyA*** *tuma kala Aoge* ?
'?' 'you' 'tomorrow' 'will come'
"Will you come tomorrow ?

5.12.1 QF Quantifiers

All quantifiers like Hindi *kama* (less), *jyAdA* (more), *bahuwa* (lots), etc. will be marked as QF.

h34. *vahAz* ***bahuta*****_QF** *loga Aye the*
'there' many' 'people' 'came' 'was'
"Many people came there".

In case these words are used in constructions like '***baHutoM ne*** *jAne se inkAra kiyA*' ('many' 'by' 'to go' 'refused'; Many refues to go) where it is functioning like a noun, it will be marked as NN (noun). Quantifiers of number will be marked as below.

5.12.2 QC Cardinals

Any word denoting a cardinal number will be tagged as QC. Penn tag set has a tag CD for cardinal numbers and they have not talked of ordinals. For example,
h35. *vahAz* ***tIna*****_QC** *loga bEThe the*
'there' 'three' 'people' 'sitting' 'were'
"Three people were sitting there"

5.12.3 QO Ordinals

Expressions denoting ordinals will be marked as QO.

h36. *mEMne kitAba* ***tIsare*****_QO** *laDake ko dI thI*
'I' 'book' 'third' 'boy' 'to' 'give' 'was'
I gave the book to the third boy"

5.12.4 CL Classifiers

The tag CL has been included to mark classifiers. Many Indian languages have a rich classifier system. "A **classifier**, in linguistics, is a word or morpheme used in some languages to classify a noun according to its meaning" (http://en.wikipedia.org/wiki/Classifier_%28linguistics%29).

For example,

Telugu : (t2) *padi* ***mandi*** *pillalu* 'ten' 'persons' 'children'

Tamil : (tm1) *pattu* ***pEr*** *mANavarakaLa*
'ten' 'person' 'students'

The words 'mandi' (Telugu) and 'per' (Tamil) are classifiers which occur with numerals with human nouns. Such expressions when occurring separately (not suffixed with the noun) will be marked as CL. Therefore :

Telugu : (t2) *padi* ***mandi*_CL** *pillalu*
'ten' 'persons' 'children'

Tamil : (tm1) *pattu* ***pEr*_CL** *mANavarakaLa*
'ten' 'person' 'students'

5.13 INTF Intensifier

This tag is not present in Penn tag set. Words like *'bahuta*', *'kama'*, etc. when intensifying adjectives or adverbs will be annotated as INTF. Example,

h37. *hEdarAbAda meM aMgUra* **bahuta_INTF** *acche milate hEM*
'HyderabAd' 'in' 'grapes' 'very' 'good' 'available' 'are'
“Very good grapes are available in Hyderabad”.

5.14 INJ Interjection

The interjections will be marked as INJ. Apart from the interjections, the affirmatives such as Hindi 'HAz'('yes') will also be tagged as INJ. Since, this is the only example of such a word, it has been clubbed under Interjections.

h38. ***arre*_INJ**, *tuma A gaye !*
'oh' 'you' 'come' 'have'
“Oh! you have come”

h39. ***hAz*_INJ**, *mEM A gayA*
'yes', 'I' 'come' 'have'
“Yes, I have come”.

5.15 NEG Negative

Negatives like Hindi 'nahIM' (not), 'nA' (no, not), etc. will be marked as NEG. For example,

h40. *vaha Aja* ***nahIM*_NEG** *A pAyegA*
'he' 'today' 'not' 'come' 'will be able'

Also, see examples (b2) and (h25) given above.
Indian languages have reiteration of NEG in certain constructions. For example,

b5. t*umi chobitA dekhbe* ?
'you' 'picture-def' 'will see' ?
“Will you see the picture ?”
b6. *nA*_NEG, *xekhabo nA*_NEG
'*n*o' 'will see (I)' 'not'
“No, I will not see (it)”

The first occurrence of *'nA'* in such constructions will also be marked as NEG.

5.16 UT Quotative

A quotative introduces a quote. Typically, it is a verb. Many Indian languages use quotatives. Given below is an example from Bengali,

b7. *she Ashbe* ***bole*** *bolechilo*
'he' 'will come' 'quotative' 'told'
“He told that he will come”.

5.17 SYM Special Symbol

All those words which cannot be classified in any of the other tags will be tagged as SYM. This tag is similar to the Penn 'SYM'. Also special symbols like $, %, etc are treated as SYM. Since the frequency of occurrence of such symbols is very less in Indian languages, no separate tag is used for such symbols.

5.18 *C Compounds (Make it XC – where X is a variable of the type of the compound of which the current word is a member of)

The issue of including a tag for marking compounds was discussed extensively. Results of algorithms using IIIT-H tag set which included NNC (part of compound nouns) and NNPC (part of proper nouns) showed that these two tags contributed substantially to the low accuracy of the tagger. Since most elements which occur as NNC or NNPC can also occur as NN and NNP, it affected the learning by the machine. So, the question was, why to include tags which contributed more to the errors ? The other aspect, however, was that while human annotators are annotating the data, they know from the context when a certain element is NNC or NN, NNPC or NNP and if marked, this information can be useful for certain applications. The argument is same as the one in favor of including a tag for proper nouns.

Another point which was discussed was that any word class can have compound forms in Indian languages (including adjectives and adverbs).

Therefore, if we decide to have a tag for showing compounds of each type, the number of tags will go very high. The final decision on this was to include a *C tag which will be realised as **catC** tag of the type of compound that the element is a part of. For example, if a certain word is part of a compound noun, it will be marked as NNC, if it is part of a compound adjective, it will be marked as JJC and so on and so forth.
Some examples are given below :

Hindi compound noun *keMdra sarakAra* (Central government) will be tagged as *keMdra_*NNC *sarakAra_*NN.

In this example, '*keMdra*' and '*sarakAra*' are both nouns which are forming a compound noun. All words except the last one, of a compound words will be marked as NNC. Thus any NNC will be always followed by another NNC or an NN. This strategy helps identify these words as one unit although they are not conjoined by a hyphen. Similarly, a compound proper noun will be marked as NNPC excluding the last one. eg. *aTala_*NNPC *bihArI_*NNPC *vAjapeyI_*NNP
The first two words, in the above example, will be tagged as NNPC and the last one will be tagged as NNP. Similar to the NNC tag for common nouns, NNPC tag helps in marking parts of a proper noun.

h41. *rAma, mohana aur shyAma ghara gaye.*
'Ram', 'Mohan' 'and' 'Shyam' 'home' 'went'
"Ram, Mohan and Shama went home".

h42. *bagIce meM* ***ranga_*JJC *biraMge_*JJ** *phUla khile the*
'garden' 'in' 'colourful' 'flowers' 'flowered' 'were'
"The garden had colorful flowers"

Titles such as **Dr., Col., Lt**. etc. which may occur before a proper noun will be tagged as **NNC**. All such titles will always be followed by a Proper Noun. In order to indicate that these are parts of proper nouns but are nonetheless nouns themselves, they will be tagged as NNC, eg. **Col._NNC** Ranjit_**NNPC** Deshmukh_**NNP**

5.19 RDP Reduplication

In this phenomenon of Indian languages, the same word is written twice for various purposes such as indicating emphasis, deriving a category from another category etc. eg. *choTe choTe* ('small' 'small'; very small), *lAla lAla* ('red' 'red'; red), *jaldI jaldI* ('quickyl' 'quickly' ; very quickly)
There are two ways in which such word sequences may be written. They can be written – (a) separated by a space or (b) separated by a hyphen.

The question to be resolved is that in case, they are written as two words (separated by space)– how should they be tagged? Earlier decision was to use the same tag for both the words. However, in this approach, the morphological character of reduplication is missed out. That is, the reduplicated item will then be treated exactly like two independent words of the same category. For example,

h43. *vaha* ***mahaMgI_*****JJ** ***mahaMgI_*****JJ** *cIjZeM kharIda lAyA*
'he' 'expensive' 'expensive' 'things' 'buy' 'bring'
“He bought **all expensive** things”.

h44. *una* ***catura_JJ buddhimAna_JJ*** *baccoM ne samasyA sulajhA lI*
'those' 'smart' 'intelligent' 'children' 'erg' 'problem' 'solved'
“Those **smart** and **intelligent** children solved the problem.

Both (h43) and (h44) have a sequence of adjectives - *mahaMgI_*JJ *mahaMgI_*JJ and *catura_JJ buddhimAna_JJ* respectively. In the first case, the sequence of two adjectives is a case of reduplication (same adjective is repeated twice to indicate the intensity of 'expensive') whereas in the second case the two adjectives refer to two different properties attributed to the following noun. Since reduplication is a highly productive process in Indian languages, it is proposed to include a new tag **RDP** for annotating reduplicatives. The first word in a reduplicative construction will be tagged by its respective lexical category and the second word will be tagged as RDP to indicate that it is a case of reduplication distinguishing it from a normal sequence such as in (h44) above. Some more examples are given underneath to make it more explicit,

h45. *vaha* ***dhIre_*****RB** ***dhIre_*****RDP** *cala rahA thA.*
'he' 'slowly' 'slowly' 'walk' 'PROG' 'was'
“He was walking (very) slowly”.

h46. *usake bAla* ***choTe_*****JJ** ***choTe_*****RDP** *the.*
'his' 'hair' 'short' 'short' 'were'
“He had (very) short hair”

h47. yaha bAta ***galI_*****NN** ***galI_*****RDP** *meM phEla gayI.*
'this' 'talk' 'lane' 'lane' 'in' 'spread' 'went'
“The word was spread in every lane”.

5.20 ECH Echo words

Indian languages have a highly productive usage of echo words such as Hindi '*cAya-vAya*' ('tea' 'echo'), where *'cAya*' is a regular lexical item of Hindi vocabulary and *'vAya*' is an echo word indicating the sense “etc” . These words, on their own, are 'nonsense' words and do not find a place in any dictionary. Thus, the gloss for '*cAya-vAya*' would be '*tea etc*'. It is proposed to add the tag **ECH** for such words.

5.21 UNK Unknown

A special tag to indicate unknown words is also included in the tag set. The annotators can use this tag to mark the words whose category they are not aware of. This tag has to be used very cautiously and sparsely, i.e., only if it is absolutely necessary.

6. Some Special Cases

This section gives the details of certtain aspects of Indian languages which need to be dealt with separately in the tagger. These are issues that cannot be handled by just changing or adding tags.

6.1 'vAlA' type constructions

'*vAlA*' is a kind of suffix used in Hindi and some other Indian languages. It conjoins with nouns (Case I, below) or verbs (Case II) to form adjectives or even nouns. It is also used as an aspectual TAM in a verbal construction (Case III).

h48. *lAla* ***kamIjZa vAlA*** *AdamI merA bhAI hE* .
'red' 'shirt' 'in' 'man' 'my' 'brother' 'is'
“The man in red shirt is my brother”.

h49. *mehanawa* ***karane vAle*** *vyakti ko inAma milegA* .
'hard work' 'doing' 'adj' 'person' 'to' 'prize' 'will get'
The person who works hard will get a prize.

These cases are elaborated below.

Case I: The suffix 'vAlA' can occurr with a noun. For example, ***lAThI vAlA (*** 'stick' 'with' **-**The one with a stick).

h50. *lAThI vAle AdamI ko bulAo*
'with stick' 'man' 'acc' 'call'
“Call the man with the stick”.

This suffix '*vAlA*' in Hindi (a) may be written separately or (b) may be attached to the preceding noun.

(a) In case it is written separately as in '*lAThI vAlA*' above, the word '*lAThI*' will be tagged as NN and the word '*vAlA*' will be tagged as PSP.

The whole expression '*lAThI vAlA*' is an adjective, in which '*lAThI*' is a noun and '*vAlA*' is a suffix which derives an adjective from a noun. Since '*lAThI*' and '*vAlA*' written separately in the above example, they have to be tagged individually. '*vAlA*' in such cases will be treated like a postposition and will be tagges as PSP.

(b) The second possibility is of '*lAThi*' and '*vAlA*' written together as '*lAThIvAlA*'. In such cases it will be treated as one word and will be marked as JJ since 'l*AThIvAlA*' is an adjective.

Case II: '*vAlA*' can also occur after a verb. Example, ***karane vAlA*** ('doing' 'one' – The one who does something)

h51. *mehanata* ***karanevAle ko*** *phala milatA hE*
'hard' 'working one' acc 'fruit' 'get' 'PRES'
The one who works hard gets the fruits".

As mentioned earlier, the suffix '*vAlA*' also joins a verb in its nominal form and makes it an adjective. In this case also, the two words may be written separately (*karane vAle*) or together (*karanevAlA*). In the former case, the two words will be marked as VM and PSP respectively (*karane*_VM *vAle*_PSP). In the latter case, being a single word (*karanevAlA*) it will be tagged as VM (*karanevAle*_VM). It is crucial to retain the 'verb' information in these case, so that at a later stage if we want to annotate its argument structure we should be able to do so (discussed earlier in the document).

Case III: 'vAlA' can also occur as part of TAM. For example,

h52. *mEM wo* ***jAne vAlA*** *hI thA.*
'I' 'particle' 'to go' 'about' 'part' 'was'
"I was about to go"

Although the word '*jAne*' has a '*vAlA*' suffix in (h52) above, the entire expression is not an adjective but is a verb having the aspectual information of 'shortly'. In this case, the sequence '*jAne vAlA*' will be marked as *jAne*_VM *vAlA*_VAUX. The alternative writing convention of writing the sequence as one word (*jAnevAlA*) is possible in this case also. Like the earlier cases, the word will be marked for the category of the content morpheme – which is verb in this case. Thus *jAnevAle* will be tagged as *jAnevAle*_VM.

Here again we stand by our policy that the tag will be decided on the basis of the part of speech and not on the basis of the category of the word in the given sentence(syntactic function). This avoids confusion at the level of manual tagging and aids machine learning as well. So the tag (VM) remain same although the function of the words is different in two different places, it is adjective in Cases I and II and verbal in Case III.

6.2 Honorifics in Indian languages

Hindi (and some other Indian languages) has particles such as '*jI*' or '*sAHaba*' etc. after proper nouns or personal pronouns. These particles are added to denote respect to the referred person. Such honorific words will be treated like particles and will be tagged **RP** like other particles.

h53. ***mantrI*_NN *jI*_RP** *sabhA meM dera se pahuMce .*

'minister' 'hon' 'meeting' 'in' 'late' 'part' 'reached'
“The minister reached late for the meeting”.

6.3 Foreign words

Presence of loan words is a fairly common phenomenon in languages. Most Indian languages have a number of loan word from English. One may also come across words from other Indian languages or Sanskrit in a given text. Such foreign words will be tagged as per the syntactic function of the word in the given context. In special cases, such as when the annotator is not sure of the category of a word, it will be tagged as **UNK.**

7. A Special Note

There may be situations, when an annotator does not feel very confident about the tag for a particular word. The annotator may then assign it different tags in different places. Inconsistency in the manual tagging can affect the learning considerably. Since this is a task which involves a number of human annotators, the methods have to be evolved to check and cross validate the human annotation. Another practical problem in annotation is that in the initial stages of annotation, the annotators need time to get familiar with the tagging scheme and the concept behind each tag. Thus they take some time before coming to a stable stage of decision making for various instances, particularly various ambiguous cases. Especially, in the initial stages, the annotators may often come across cases where their confidence level may not be very high. They may feel the need of some clarifications for these cases. Since the task of annotation has to go on and immediate clarification may not be possible, the annotators may be forced to take decisions and mark a case as they consider appropriate at that point of time. Over a period of time, with better understanding of the tags and tagging scheme, they may reach a stable stage. However, by then they may already have tagged a given case differently in different places thus introducing inconsistency in the annotated corpus. At a later stage, it will be difficult to go back to all the cases that have been annotated by then and correct them. So the chances are that the annotators may proceed with the revised decision and leave the earlier annotation as such. This will introduce inconsistencies in the annotated corpus.

To control such a situation, it is decided to provide a way by which the annotators can initially mark the uncertainty of their decision so that they can easily extract these cases easily and take them up for discussions and clarifications.

This 'uncertainty' will be annotated as follows :

a) The annotators first mark such a case with a tag that they consider appropriate at the time of annotation.

b) Along with the chosen tag, they also put a question mark (?) against that tag. The question mark will indicate that this case is not yet resolved and will be finalized after clarification or discussion.
c) All the cases with a question mark can be later taken out and placed for discussion. An annotator will be responsible for bringing such cases for discussion and once the cases are resolved, the annotator will go back and correct the tag. In case the tag assigned by the annotator initially itself is correct, the annotator will remove the question mark against it.

This is a purely temporary measure and the data finally submitted by an annotator should not have any words having a question mark.

8. Chunk Tags Chosen for the Current Scheme

This section deals with the chunk tags. Not many of the issues discussed above hold for defining the chunk tags. Various points which have been deliberated upon in relation to chunking scheme are :

1. Definition of a chunk
2. Chunk Types
3. Some Special Cases
4. Annotation method/procedure

8.1 Definition of a chunk

Following issues related to the definition of a chunk were discussed :
What constitutes a 'chunk' ?

A typical chunk consists of a single content word surrounded by a constellation of function words (Abney,1991). Chunks are normally taken to be a 'correlated group of words'.

The next issue, however, is - How to define the boundaries of these 'correlated word groups' for our purpose?

For example, which case in the following pairs should be grouped as a chunk ?

((*xillI meM*)) OR ((*xillI*)) *meM*
'Delhi' 'in' *'Delhi' 'in'*
((*rAjA kA betA*)) OR ((*rAjA kA*)) ((*betA*))
'king' 'of' 'son' 'king 'of' 'son'
((*rAjA ke bete kI paxnI*)) OR ((*rAjA ke*)) ((*bete k*I)) ((*paxnI*))
'king' 'of' 'son' 'of' 'wife' 'king' 'of' 'son' 'of' 'wife'

Following definition of a 'chunk' was evolved through discussion :

“A minimal (non recursive) phrase(partial structure) consisting of correlated, inseparable words/entities, such that the intra-chunk dependencies are not distorted”. Each chunk type discussed and the decided upon is described below .

8.2. Chunk Types

Based on the above definition of chunk, issues related to various chunk types were discussed. A chunk would contain a 'head' and its modifiers.

8.2.1 NP Noun Chunk

Noun Chunks will be given the tag NP and include non-recursive noun phrases and postpositional phrases. The head of a noun chunk would be a noun. Specifiers will form the left side boundary for a noun chunk and the vibhakti or head noun will mark the right hand boundary for it. Descriptive adjective/s modifying the noun will be part of the noun chunk. The particle which anchors to the head noun in a noun chunk will also be grouped within the chunk. If it occurs after the noun or vibhakti, it will make the right boundary of the chunk. Some example noun chunks are :

((*bacce*_NN))_NP, ((*kucha*_QF *bacce*_NN))_NP,
'children' 'some' 'children'
((*kucha*_QF *acche*_JJ *bacce*_NN))_NP, ((*Dibbe*_NN *meM*_PSP))_NP,
'some' 'good' 'children' 'box' 'in'
((eka_QC *kAlA*__JJ *ghoDZA*_NN))_NP ,
'one' 'black' 'horse'
((*yaha*_DEM *nayI*_JJ *kitAba*_NN))_NP,
'this' 'new' 'book'
((**isa**_DEM *nayI*_JJ *kitAba*_NN *meM*_PREP))_NP,
'this' 'new' 'book' 'in'
((*isa*_DEM *nayI*_JJ *kitAba*_NN *meM*_PSP *bhI*_RP))_NP
'this' 'new' 'book' 'in' 'also'

The issue of genitive marker and its grouping with the nouns that it relates to was discussed in detail. For example, the noun phrase '*rAma kA beTA*' contains two nouns '*rAma*' and '*beTA*'. The two nouns are related to each other by the vibhakti '*kA*'. The issue is whether to chunk the two nouns separately or together? Linguistically, '*beTA*' is the head of the phrase “*rAma kA beTA*”. '*rAma*' is related to '*beTA*' by a genitive relation which is expressed through the vibhakti '*kA*'. Going by our definition of a 'chunk' we should break '*rAma kA beTA*' into two chunks (((*rAma kA*))_NP, ((*beTA*))_NP) by breaking '*rAma kA*' at 'kA' vibhakti . Moreover, if we chunk 'rAma kA beTA' as one chunk, linguistically, we will end up with a recursive noun phrase as a single chunk ((((*rAma kA)) beTA*)) which also is against our definition of a chunk.

Therefore, it was decided that the genetive markers will be chunked along with the preceding noun. Thus, the noun group 'rAma kA beTA' would be chunked into two chunks.

h54. **((*rAma kA*))NP** ((*beTA*))NP acchA hE "Ram's son is good"
h55. ((*kitAba*))NP **((*rAma kI*))NP** hE "The book belongs to Ram"

For the noun groups such as "*usakA beTA*" it was decided that they should be chunked together.

8.2.2 Verb Chunks

The verb chunks would be of several types. A verb group will include the main verb and its auxiliaries, if any. Following are some examples of verb chunks from Hindi,

((*khAyA*)), ((*khA rahA hE*)), ((*khA sakawe hEM*))
'ate' 'eat' 'PROG' 'is' 'eat' 'can' 'PRES'

The types of verb chunks and their tags are described below.

8.2.2.1 VGF Finite Verb Chunk

As mentioned in 5.4 above, a verb group sequence (V VAUX VAUX . .) contains a main verb and its auxiliaries. The group itself can be finite or non-finite. In case of it being finite, the main verb in such a sequence may not be finite. The finiteness is known by the auxiliaries. Therefore, it is decided to mark the finiteness of the verb at the chunk level. Thus, any verb group which is finite will be tagged as **VGF.** For example,

h56. *mEMne ghara para khAnA ((khAyA*_VM*))*_**VGF**
'I erg' 'home' 'at' 'meal' 'ate'

h57. *vaha cAvala ((khA*_VM *rahA*_VAUX *hE*_VAUX*))*_**VGF**
'he' 'rice' 'eat' 'PROG' 'is'

8.2.2.2 VGNF Non-finite Verb Chunk

A non-finite verb chunk will be tagged as **VGNF.** For example,

h15a) *seba* **((*khAtA*_VM *huA*_VAUX))_VGNF** *laDZakA jA rahA thA*
'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

h16a) *mEMne* **((*khAte – khAte*_VM))_VGNF** *ghode ko dekhA*
'I erg' 'while eating' 'horse' acc 'saw'
h17a) *mEMne ghAsa* **((*khAte*_VM *hue*_VAUX))_VGNF** *ghoDe ko dekhA*
'I erg' 'grass' 'eating' 'PROG' 'horse' acc 'saw'

The IIIT-H tagset had initially included three tags for the non-finite verbal forms. Unlike Penn tagset, all non finite verbs, which are used as adjectives, were marked as VJJ at the POS level. Similarly, to mark adverbial non-finite verbs, the POS tagset had VRB tag. A tag VNN was included to mark the nominalized verbs.

However, during the discussions IL standards, it was pointed out that inclusion of too many finer tags hampers machine learning. Moreover, the marking is based on syntactic information, which we should avoid at the POS level, unless it is contributing to further processing in a substantial way. On the other hand, it is important to mark finite non-finite distinction in a verbal expression as it is a crucial information and is also easy to learn. As discussed under 5.4 above, it was decided to mark this distinction at the chunk level, rather than at the POS level. Therefore, the tag VGNF has been included to mark non-finite adverbial and adjectival verb chunk.

8.2.2.3 VGINF Infinitival Verb Chunk

This tag is to mark the infinitival verb form. In Hindi, both, gerunds and infinitive forms of the verb end with a *-nA* suffix. Since both behave functionally in a similar manner, the distinction is not very clear. However, languages such as Bangla etc have two different forms for the two types. Examples from Bangla are given below.

b8. *Borabela **((snAna karA))_*****VGNN** *SorIrera pokze BAlo*
'Morning' 'bath' 'do-verbal noun' 'health-gen' 'for' 'good'
‘Taking bath in the early morning is good for health”

b9. *bindu Borabela ((**snAna karawe))**_***VGINF** *BAlobAse*
'Bindu' 'morning' 'bath' 'take-inf' 'love-3pr'
“Bindu likes to take bath in the early morning”

In Bangla, the gerund form takes the suffix *–A / -Ano*, while the infinitive marker is *–we*. The syntactic distribution of these two forms of verbs is different. For example, the gerund form is allowed in the context of the word *darakAra* “necessary” while the infinitive form is not, as exemplified below:

b10 *Borabela ((snAna karA))_***VGNN** *darakAra*
'Morning' 'bath' 'do-verbal noun' 'necessary'
“It is necessary to take bath in the early morning”

b11. **Borabela ((snAna karawe))_***VGINF** *darakAra*

Based on the above evidence from Bangla, the tag *VGINF* has been included to mark a verb chunk.

8.2.2.4 VGNN Gerunds

A verb chunk having a gerund will be annotated as **VGNN**. For example,

h18a. *sharAba ((****pInA*****_VM))_VGNN** *sehata ke liye hAnikAraka hE.*
'liquor' 'drinking' 'heath' 'for' 'harmful' 'is'
“Drinking (liquor) is bad for health”

h19a. *mujhe rAta meM ((****khAnA*****_VM))_VGNN** *acchA lagatA hai*
'to me' 'night' 'in' 'eating' 'good' 'appeals'
“I like eating at night”

h20a. ((***sunane*****_VM** ***meM*****_PSP***))*_**VGNN** *saba kuccha acchA lagatA hE*
'listening' 'in' 'all' 'things' 'good' 'appeal' 'is'

8.2.3 JJP Adjectival Chunk

An adjectival chunk will be tagged as **JJP**. This chunk will consist of all adjectival chunks including the predicative adjectives. However, adjectives appearing before a noun will be grouped together with the noun chunk. A JJP will consist of phrases like

h58. *vaha laDaZkI hE*((*suMdara*_**JJ** *sI*_RP))_**JJP**
'she' 'girl' 'is' 'beautiful' 'kind of'

h59. *hAthI AyA* ((*moTA*_*C *tagadA*_**JJ**))_**JJP**
'elephant' 'came' 'fat' 'powerful'

h60. *vaha laDakI ((bahuta*_INTF *sundara*_**JJ**))_**JJP** *hE*
'she' 'girl' 'very' 'beautiful' 'is'

Cases such as (h61) below will not have a separate JJP chunk. In such cases, the adjectives will be grouped together with the noun they modify. Thus forming a NP chunk.

h61. ((*kAle*_**JJ** *ghane*_**JJ** *laMbe*_**JJ** *bAla*_**NN**))_**NP**
'black' 'thick' 'long' 'hair'

8.2.3.1 Some special cases

Following examples from Hindi present a

h62. *xillI meM* ***rahanevAlA*** *merA BAI kala A rahA hE .*
'Delhi' 'in' 'staying' 'my' 'brother' 'tomorrwo' 'come' 'PROG' 'is'
“My brother who stays in Delhi is coming tomorrow”.
h63. *usane Tebala para* ***rakhA huA*** *seba khAyA.*
'(s)he erg' 'table' 'on' 'kept' 'apple' 'ate'

“He ate the apple kept on the table”.

In (h62) above '*rahanevAlA*' is an adjectival participle. But we do NOT mark it as JJP. Instead, it will be marked as a **VGNF**. The decision to tag it as a VGNF is based on the fact that such adjectival participles are derived from a verb can have their arguments. This information is useful for processing at the syntactic level. Thus, '*rahanevAlA*' in (h62) will be annotated as follows:

h62a. *xillI meM* ((***rahanevAlA*_VM*)*_VGNF** *merA BAI kala A rahA hE* .

Similarly, in (h63) above, the chunk '*rakhA huA*' is an adjective but will also be marked as a VGNF since this also derived from a verb and chunks like '*Tebala pra*' etc are its arguments. So the chunk name will be **VGNF** and the POS tag will be **VM** which might be followed by an auxiliary verb tagged as **VAUX**. (h63a) shows how '*rakhA huA*' will be annotated :

h63a. *usane Tebala para ((***rakhA*_VM huA*_VAUX*))*_VGNF***seba khAyA*.

8.2.4 RBP Adverb Chunk

This chunk name is again in accordance with the tags used for POS tagging. This chunk will include all pure adverbial phrases.

h64. *vaha* ((*dhIre-dhIre*_**RB**))_**RBP** *cala rahA thA.*
'he' 'slwoly' 'walk' 'PROG' 'was'
“He was walking slowly”

Now consider the following examples:

h65. *vaha* ***dagamagAte hue*** *cala rahA thA* .
'he' ' 'walk' 'PROG' 'was'
“He was walking

h66. *vaha khAnA* ***khAkara*** *ghara gayA* .
'he' 'meal' 'after eating' 'home' 'went'
“He went home after eating his meal”

In the above examples, '*dagamagAte hue*' and 'k*hAkara*' are non finite forms of verbs used as adverbs. Similar to adjectival participles these will also be chunked as **VGNF** and not as **RBP**. The reason for this is that we need to preserve the information that these are underlying verbs. This will be a crucial information at the level of dependency marking where the arguments of these verbs will also be marked.

((isa_PRP nayI_JJ kitAba_NN meM_PSP bhI_RP))_NP
'this' 'new' 'book' 'in' 'also'

8.2.5 NEGP Negatives

(i) In case a negative particle occurs around a verb, it is to be grouped within verb group. For example,

h67. *mEM kala dillI ((**nahIM**_*NEG ***jA***_VM ***rahI***_VAUX*))_*VGF
"I" "tomorrow" "Delhi" "not" "go" "Cont"

h68. ((***binA***_NEG ***bole***_VM))_VGNF *kAma ((**nahIM**_*NEG ***calatA***_VM*))_*VGF
"without" "saying" "work" "not" "happen"

However ,

h69. **binA** kucha **bole** kAma **nahIM calatA**
"without" "something" "saying" "work" "not" "happen"

In the above sentence, the noun "*kucha*" is coming between the negative "*binA*' and verb "*bole*". Here, it is not possible to group the negative and the verb as one chunk. At the same time, "*binA*" cannot be grouped within an NP chunk, as functionally, it is negating the verb and not the noun. To handle such cases an additional **NEGP** chunk is introduced. If a negative occurs away from the verb chunk, the negative will be chunked by itself and chunk will be tagged as NEGP. Thus,

h69a. **((*binA*))_NEGP** **((***kucha*))_NP **((*bole*))_VG** **((***kAma*))_NP ((*nahIM calatA*)**)_VG**

8.2.6 CCP Conjuncts

Conjuncts are functional units information about which is required to build the larger structures. Take the following examples of cunjunct usages :

h70. (*rAma kitAba paDha rahA thA)* ***Ora*** *(mohana Tennisa khela rahA thA).*
"Ram was reading a book **and** Mohan was playing tennis"

h71. (*rAma ne batAyA)* ***ki*** *(usakI kitAba acchI hE).*
"Ram said **that** his book is good"

h72. (*rAma)* **Ora** *(mohana) Tennisa khela rahe the.*
"Ram **and** Mohan were playing tennis".

h73. (*merA bhAI rAma)* ***Ora*** *(usakA dosta mohana) Tennisa khela rahe the.*
"My brother Ram **and** his friend Mohan were playing tennis".

h74 . *rAma (saphZeda kapade)* ***Ora*** *(nIle jute) pahane thA.*
"Ram was wearing white clothes and blue shoes".

h75. *rAma eka (halkI)* ***Ora*** *(nIlI) bOla lAyA.*
"ram brought a light **and** blue ball".

The sentences above have various types of conjoined structures. To represent these conjoined structures, it is decided to form separate chunks for conjuncts and the elements a conjunct conjoins. Thus (h70) and (h71) above will be chunked as (h70a) and (h71a) given below,

h70a. ((*rAma*))_NP ((*kitAba*))_NP ((*paDha rahA thA*))_VG **((*Ora*))CCP** ((*mohana*))_NP ((*Tennisa*))_NP ((*khela rahA thA*))_VG.

h71a. ((*rAma ne*))_NP ((*batAyA*))_NP **((*ki*))_CCP** ((*usakI*))_NP ((*kitAba*))_NP ((*acchI*))_JJP ((*hE*))_VG.

Expression '*rAma Ora mohana*' in example (h72) is a complex NP. Though complex, the expression can be annotated as a single NP chunk as functionally it is the subject of the verb 'play'. However, example (h73) presents a case where it would be better to form three independent chunks for the complex subject NP. Though the conjunct '***Ora***' is conjoining '*rAma*' and '*mohana*', both '*rAma*' and '*mohana*' have their respective modifiers. To make it explicit, it is better to treat them as two independent NP chunks conjoined by a CCP.

h73a. ((*merA bhAI rAma*)) **((*aura*))_CCP** ((*usakA dosta mohana*))_NP ((*Tennisa*))_NP ((*khela rahe the*))_VG.

Following this, the subject NP of (h72) would also be annotated similarly. Therefore,

h72a. ((*rAma*))_NP **((*aura*))_CCP** ((*mohana*))_NP ((*Tennisa)*)_NP ((*khela rahe the*))_VG.

The annotation for cases such as (h74) and (h75) would be as follows :

h74a. ((*rAma*))_NP ((*safeda kapade*))_NP **((*aura*))_CCP** ((*nIle jute*))_NP ((*pahane thA*))_VG.

h75a. ((*rAma*))_NP ((*eka*))_JJP ((*halkI*))_JJP **((*aura*))_CCP** ((*nIlI*))_JJP ((*bOla*))_NP ((*lAyA*))_VG

Thus the decision for conjuncts is - the conjoined entities will be broken into separate chunks. eg. ((*rAma*))_NP ***((Ora))_*****CCP** ((*SyAma*))_NP

8.2.7 FRAGP Chunk Fragments

Some times certain fragments of chunks are separated from the chunks to which they belong. For example :

h76. ***rAma*** (*jo merA baDZA bhAI hE*) ***ne*** *kahA ...*
'Ram' 'who' 'my' 'elder' 'brother' 'is' 'erg' 'said'

In the above example, vibhakti '*ne*', which is a case marker of the noun '*rAma*', is separated from it by an intervening clause. Syntactically, '*ne*' is a part of the noun chunk '*rAma ne*'. However, at times it can be written separately. The following was decided for such fragments :

(i) There will be a separate chunk for the vibhakti in constructions where it gets separated from the noun it would normally be grouped with. This chunk can have more than one entity within it.

h77. ((***rAma***))**_NP,** *mere dillI vAle bhAI,* **((*ne*))_FRAGP** *kahA*
'Ram' 'my' 'Delhi' 'from' 'brother' 'erg' 'said'

(ii) If the entities embedded between the noun and it's vibhakti are a series of nouns the entire group will be chunked as a single noun chunk.

h78. **((***isa* ***'upanyAsa samrATa'*** *Sabda kA***))_NP**
'this' 'Novel' 'King' 'word' 'of'

8.2.8 BLK Miscellaneous entities

Entities such as interjections and discourse markers that cannot fall into any of the above mentioned chunks will be kept within a separate chunk.
eg. ((*oh*_INJ))**_BLK,** ((*arre*_INJ))**_BLK**

8.3 Some Special Cases

Apart from the above, some special cases related to certain lexical types are discussed below.

8.3.1 Conjunct Verbs

The issue whether to treat the noun/adjective which is part of a conjunct verb differently by marking it with a special tag (NVB/JVB) or to treat it as a noun like any other noun at the POS level was deliberated on.
The question was based on the following observations :

a) NVB/JVB , as part of conjunct verbs, are most often not recognized by the learning algorithms.

b) Having NVB at the POS level is based on syntactic considerations. Therefore, do we really need to go for it ? Instead, at the POS level we mark the noun as a noun and leave the decision of marking a conjunct verb as single unit for a later level.

c) Moreover, since the noun, which is part of a conjunct verb (Kriyamula), can occur away from its 'verbaliser', it becomes difficult to differentiate it from a 'noun' which may be an argument of the verb. This also creates problem for chunking of the verb group. The two components of the chunk have to be separately marked and have to be joined at the syntactic level.

d) If NVB is marked at the POS level, a natural consequence would be to group it with its verbaliser as a VG chunk. In fact, that is the purpose of identifying it as different from a noun. However, sometimes one comes across expressions such as '*mEMne unase eka* ***prashna kiyA***' (I posed a question to him). In this sentence, '***eka***' is a modifier of '*prashna*'. '*prashna karana*' is recognized as a conjunct verb in Hindi by most Hindi speakers. Following example shows the problem of grouping '*praSna karanA*' as a single VG:

POS : *mEMne*_PRP *unase*_PRP *eka*_QC *prashna*_NVB *kiyA*_VM
Chunk : ((*meMne*_PRP))_NP ((*unase*_PRP))_NP ((*eka*_QC))_JJP ((*prashna*_NVB *kiyA*_VM))_VGF

Once "*praSna karanA*" are grouped together as a chunk, it will be difficult to show the relation between '*eka*' and 'prashna' subsequently.

Thus, an alternative was proposed wherein, the noun of the conjunct verb is tagged as NN at the POS level which is accordance with the decision to tag the lexical item based on its lexical category. Thereafter, the noun is grouped with its preceding adjectival modifiers as an NP chunk. The only problem in this approach is that the information of a noun verb sequence being a conjunct verb is not captured till the chunk level and the noun of the conjunct verb is separated from its verbaliser. However, the approach has following advantages :

1) At the POS level, the word is tagged for its grammatical category and not for its syntactic function. This eases the decision making at the POS level. And marking the information, that the conjunct verbs which are composed of two words form one lexeme semantically, is postponed to a later level.

2) It allows us to show the modifier-modified relation between an adjective such as '*eka*' in the above example with its modified noun '*praSna*'.

3) Since the information of a noun verb sequence being a 'kriyamula' is crucial at the syntactic level, it will be captured at that level by marking the relation between the 'noun' and its verbaliser by an appropriate tag. Therefore, the decision is :

The noun/adjective and verb (internal components of a conjunct verb) will be chunked separately.
eg. *prashna karanA* - **((*prashna*))NP ((*kiyA*))VG**
ucita kiyA - **((*ucita*))JJP ((*kiyA*))VG**

8.3.2 Particles

Regarding the particles, it was decided that the particles will be chunked with the same chunk as the anchor word they occur with. Thus,
eg. ((*rAma ne* ***bhI***))**_NP, ((*mEM wo*))_NP,**

'Ram' 'erg' 'also' 'I' 'emph'

8.3.3 Quantifiers

The issue of chunking quantifiers was discussed in great details. Numbers can occur (a) as noun modifiers before a noun (***haZaroM*** *ladakoM ne – 'thousands' 'boys' 'erg'*) or (b) can occur without a noun (***hazAroM*** *ne – 'thousands' 'erg'*) with a nominal inflection. The issue of whether to treat the quantifiers of the type (b) as nouns was discussed. The issue is whether (b) is a case of an ellipsis of the noun after a number or whether it is the number itself which is the noun. If the latter has to be followed then the POS tag for quantifiers in such cases should be NN. Following decisions were taken :

(i) A 'QC' or 'QO' occuring with a noun will be part of the noun chunk.

h79. ((***hazAroM*_QC** *logoM_NN* *ne*_PSP))_NP *yaha driSya dekhA*
'thousands' 'people' 'erg' 'this' 'scene' 'watched'
“Thousands of people watched this scene”.

h80. ((***dUsare*_QO** *ladake*_NN *ne*_PSP))_NP *isa samasyA ko sulajhA diyA*
'second' 'boy' 'erg' 'this' 'problem' acc 'solve' 'did'
“The second boy solved this problem”.

(ii) All categories occurring without a noun, with nominal inflections (overt or otherwise) will be tagged as noun.

h81. ((***hazAroM_NN*** *ne*_PSP))_NP *yaha driSya dekhA*
'thousands' 'erg' 'this' 'scene' 'watched'
“Thousands watched this scene.

h82. **((*mote*_NN *ne*_PSP))_NP ((chote_NN ko_PSP))_NP ((***mArA***))_**VGF
'fat' 'erg' 'small' 'to' 'killed'

8.3.4 Punctuations

All punctuations, with an exception of sentence boundary markers and clausal conjuncts, will be included in the preceding chunk. For example

h83. ((*usane*_PRP))_NP ((***kahA*_VM –_SYM)_**VGF
'He erg' 'said'
((**"_SYM *yaha*_PRP**))_NP ((*Thika*_JJ))_JJP ((***hE*_VM "_SYM**))_VGF
'this' 'proper' 'is'
“He said, “this is not right” “.

h84. *rAma AyA* **((,_SYM))_CCP** *mohana gayA*
'Ram' 'came' , 'Mohan' 'went'
“Ram came and Mohan left”.

Punctuations such as (a) hyphens and (b) quote marks will be taken care of by the tokenizer.

(a) Hyphens: Identified to be of two types:-

- Without space on either sides, as in the case of compound nouns
  eg. *mAtA-pitA(mother-father)*

- With spaces, as in the case of

  h85. *rAma ne kahA – yaha thIk hE*
  'Ram' 'erg 'said' – 'this' 'proper' 'is'

(b) Quote marks (single and double both) : Identified to be of two types:-
(i) opening
(ii) closing

9. Annotation Procedure

To maintain consistency in the data format and the annotation, it was decided to use 'Sanchay', a facility developed at IIIT, Hyderabad for the annotation task.

10. Conclusion

The annotation standards for POS tagging and chunking for Indian languages include 26 tags for POS (Table-1 in Appendix) and 11 chunk tags (Table-2 in Appendix. The tags are decided on coarse linguistic information with an idea to expand it to finer knowledge if required.

11. References

Steven Abney. Parsing by Chunks. In: Robert Berwick, Steven Abney and Carol Tenny (eds.), *Principle-Based Parsing.* Kluwer Academic Publishers, Dordrecht. 1991.

Following participated in the meetings/discussions :

IIIT, Hyderabad – Rajeev Sangal, Dipti M Sharma, Soma Paul, Lakshmi Bai, Research students of LTRC, IIIT, Hyderabad.
University of Hyderabad, Hyderabad – Amba Kulkarni, G. Uma Maheshwar Rao, Rahmat Yousufzai
IIT, Bombay – Pushpak Bhattacharya, Om Damani, Rajat Kumar Mohanty, Sushant S Develkar, Pranesh, Maneesh
IIT, Khadagapur – Sudeshna Sarkar, Anupan Basu, Pratyush Banerjee, Sandipan Dandapat
CDAC, Pune – Saurabh Singhal, Abhishek Gupta, Ritu Bara, Mahendra Pandey
CDAC, Noida – Vijay Kumar, K K Arora
IIIT, Allahabad – Ratna Sanyal
Tamil University, – S. Rajendran
AUKBC, Chennai – Sobha Nair

Jadhavepur University – Shivaji Bandopadhyay

12. Acknowledgements

Ms Pranjali Karade prepared the initial document describing IIIT-H tagging scheme which has been an immense help in preparing the current document. Thanks to Soma Paul and Vasudhara Sarkar for providing Bangla examples given in the text.

13. Appendix

13.1. POS Tag Set for Indian Languages (Nov 2006, IIIT Hyderabad)

Sl No.	Category	Tag name	Example
1.1	Noun	NN	
1.2	NLoc	NST	
2.	Proper Noun	NNP	
3.1	Pronoun	PRP	
3.2	Demonstrative	DEM	
4	Verb-finite	VM	
5	Verb Aux	VAUX	
6	Adjective	JJ	
7	Adverb	RB	*Only manner adverb
8	Post position	PSP	
9	Particles	RP	bhI, to, hI, jI, hA.N, na,
10	Conjuncts	CC	bole (Bangla)
11	Question Words	WQ	
12.1	Quantifiers	QF	bahut, tho.DA, kam (Hindi)
12.2	Cardinal	QC	
12.3	Ordinal	QO	
12.4	Classifier	CL	
13	Intensifier	INTF	
14	Interjection	INJ	
15	Negation	NEG	
16	Quotative	UT	ani (Telugu), endru (Tamil), bole/mAne (Bangla), mhaNaje (Marathi), mAne (Hindi)
17	Sym	SYM	
18	Compounds	*C	
19	Reduplicative	RDP	
20	Echo	ECH	
21	Unknown	UNK	

It was decided that for foreign/unknown words that the POS tagger may give a tag ”UNK”

13.2. Chunk Tag Set for Indian Languages

Sl. No	***Chunk Type***	***Tag Name***	***Example***
1	Noun Chunk	NP	*Hindi: ((merA nayA* **ghara))_NP** *“my new house”*
2.1	Finite Verb Chunk	VGF	*Hindi: mEMne ghara para khAnA ((khAyA*_VM*))*_**VGF**
2.2	Non-finite Verb Chunk	VGNF	*Hindi: mEMne* ***((khAte – khAte*_VM*))*_VGNF** *ghode ko dekhA*
2.3	Infinitival Verb Chunk	VGINF	*Bangla : bindu Borabela* ***((snAna karawe))_*****VGINF** *BAlobAse*
2.4	Verb Chunk (Gerund)	VGNN	*Hindi: mujhe rAta meM ((***khAnA*_VM))_VGNN** *acchA lagatA hai*
3	Adjectival Chunk	JJP	*Hindi: vaha laDaZkI hE*((*suMdara*_**JJ** *sI*_RP))_**JJP**
4	Adverb Chunk	RBP	*Hindi : vaha* ((*dhIre-dhIre*_**RB**))_**RBP** *cala rahA thA*
5	Chunk for Negatives	NEGP	*Hindi:* **((*binA*))_NEGP** **((***kucha*))_NP **((*bole*))_VG** **((***kAma*))_NP ((*nahIM calatA***))_VG**
6	Conjuncts	CCP	Hindi: ((*rAma*))_NP ***((Ora))_*****CCP** ((*SyAma*))_NP
7	Chunk Fragments	FRAGP	Hindi; ***rAma*** (*jo merA baDZA bhAI hE*) ***ne*** *kahA...*
8	Miscellaneous	BLK	

Created By : Dipti M Sharma <dipti@iiit.ac.in>
Last Revision By : Dipti Misra Sharma
Creation Date : 30-11-2006
Revision Dates : 15-12-2006