

**Guidelines for Annotating Urdu TreeBank (version – 1.0)**  
Akshara Barati, Dipti Misra Sharma, Anjum Parveen, Pervez Ahmed  
Language Technologies Research Center  
IIIT, Hyderabad, India  
[dipti@iiit.ac.in](mailto:dipti@iiit.ac.in), {anjum.parveen, pervez.ahmed}@[research.iiit.ac.in](mailto:research.iiit.ac.in)

## **Contents**

### **1. Background**

### **2. The Task**

### **3. PART – 1A**

- 3.1 Grammatical Model
- 3.2 The Scheme
  - 3.2.1 Treebank Representation (SSF)
  - 3.2.2 Naming conventions
  - 3.2.3 Relations and Tag labels
- 3.3 Corpora

### **4. PART – 1B**

- 4.1 How to mark various relations?
- 4.2 How to mark elided elements?
- 4.3 How to mark shared arguments?

### **5. Some additional attributes**

### **6. PART – 2 : Urdu Example Constructions (to be included from the corpus)**

- 6.1 Simple Transitives
- 6.2 Unergatives
- 6.3 Unaccusatives
- 6.4 Dative Subject constructions (to be included)
- 6.5 Ditransitives
- 6.6 Existentials
- 6.7 Copular constructions
- 6.8 Causatives
- 6.9 Relative clauses (covered under relations, to be included here)
- 6.10 Participles (covered under relations, to be included here)
- 6.11 Complement clauses (covered under relations, to be included here)

### **7. Conclusion**

### **8. References**

### **1. Background**

A major bottleneck in developing various natural language applications for Indian languages is the unavailability of appropriate language resources. For any NLP application, certain linguistic knowledge is required. This knowledge can be prepared in the form of dictionaries, grammars, word-formation rules etc. An alternative approach is to annotate linguistic knowledge in electronic texts. The annotated texts can be used for machine learning, developing these resources by extracting the knowledge etc. Penn Treebank for English (Marcus et al., 1993), Prague Dependency Tree bank for Czech (Hajicova, 1998) etc. are some of the efforts in this direction.

The idea of developing such a resource for Indian languages was first decided to be taken up at the "Workshop on Lexical Resources for Natural Language Processing", 5 - 8 Jan 2001, held at IIIT Hyderabad. The task was named as AnnCorra, shortened for

"Annotated Corpora".

For achieving this, certain standards had to be drawn in terms of selecting a grammatical model and developing tagging schemes for the three levels of sentential analysis, POS tagging, chunking and syntactic parsing. Since Indian languages are morphologically richer, they allow the order of the words to be more flexible. This also implies that the information at the morphological level can be crucial for sentence analysis. Hence, coming up with standards for morph feature representations for various Indian languages also becomes critical. The standards for POS tagging, Chunking and Morph feature representation were arrived at in the project "IL-ILMT System". In this project nine language pairs were taken for developing bi-directional MT systems. The project is being carried out in a consortium mode and is funded by DIT, Government of India. For defining the standards for the above, several workshops were conducted with participation from major NLP groups working on the nine languages undertaken in the project.

The natural next step after POS tagging, chunking and morph analysis is sentence level parsing. Thus, it was decided to work out a scheme for annotating tree bank for Hindi. Hindi was chosen as an example language. The theoretical model that has been adopted for the sentence analysis is Panini's grammatical model which provides a level of syntactico-semantic analysis.

This document, a guidelines on dependency annotation of Urdu has two Sections. Section-1 contains a description of the grammatical model and the details of the tagging scheme. Section-2 contains examples of certain typical constructions of Hindi-Urdu and their analysis in Paninian dependency model.

## 2. The Task

The task is to develop a dependency Treebank for Urdu. As part of the task, it is decided to annotate the corpora for the following linguistic information

1. Relevant morph features for the token in the context (lexical level)
2. POS tag (lexical level)
3. Chunk (phrasal level (without distorting the internal dependencies))
4. Dependencies (sentential level – syntactico-semantic)
5. Shared and missing arguments
6. Sentence type
7. Voice type
8. Conference in specific cases

The task can be better understood with the help of an illustration. Consider the example below:

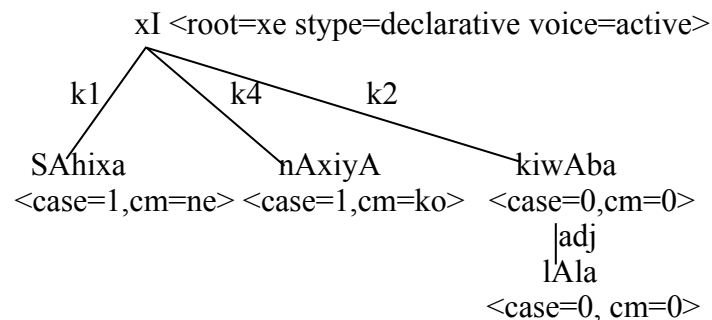
شاحد نے نادیا کو لال کتاب دی

*SAhixa ne nAxiA ko lAla kiwAba xI*

Shahid erg Nadiya acc red book gave

‘Shahid gave a red book to Nadiya.’

It would be marked as follows:



**Figure 1**

The markings here represent that *SAhixa* is the 'karta' (doer – marked as k1) of the verb *xI* 'gave', *nAxiyA* is 'sampradana' (recipient – marked as k4) and *lAla kiwAba* 'red book' is the 'karma' (locus of result of the action-k2) of the verb. Within the angular brackets under each node is provided the relevant morphological information. The root node, apart from the morphological information, also has the sentence type and voice type marked.

The main task, therefore, is to explicitly mark the relationships (arc labels) between various components of a sentence. This obviously requires a grammatical model basing which the dependency relations can be annotated.

### 3. PART 1-A

This has a description of the grammatical model used in designing the tagging scheme and the details of the tagging scheme. .

#### 3.1 Grammatical Model

Paninian grammatical model has been chosen here for the sentence analysis, hence for the tag names as well. As mentioned above, the model offers a syntactico-semantic level of linguistic knowledge which has been adopted for the Hindi treebanking. Preference for this model is based on:

- The model, not only offers a mechanism for SYNTACTIC analysis, but also incorporates the SEMANTIC information (dependency analysis).
- Indian languages have a relatively free word order, hence a dependency grammar based approach would be better suited for sentence analysis.

The Paninian grammatical model treats a sentence as a series of modifier – modified elements starting from a primary modified (generally a finite verb) . The objective of the grammarian, according to this framework, is to extract meaning from a sentence as spoken by a lay person. It works with the assumption that language is used for communication. The meaning in a sentence is encoded, not only in the lexical items, but also in the relations between words. Thus every word in a sentence has twofold role towards composing the larger meaning; (i) the concept it represents and (ii) the participatory role it plays in the sentence in relation to the other words. Most often

these roles are expressed through some explicit markers such as nominal inflections, verbal inflections etc. This implies that certain linguistic cues are explicitly available in a sentence using which one can extract the meaning from a sentence.

Morphologically rich languages such as Sanskrit (classical Indian language), Telugu, Tamil etc (some of the modern Indian languages) have the grammatical information in the words themselves. However, for languages such as Hindi-Urdu, one has to go beyond lexical items and use postpositions (for case marking) and auxiliaries (for tense, aspect, modalities) for this purpose. The scheme is then designed to meet the parsing requirements, extracting meaning from a sentence by applying important concepts in Paninian framework for achieving this.

The grammatical relations which have been considered here are of two types; (1) karaka, and (2) Relations other than karakas.

Karaka, according to Patanjali, is the one which performs an action (karotiiti kaarakam). In other words, 'karakas' are the roles of various participants in an action. An action in a sentence is denoted through a verb. For a noun to hold a karaka relation with a verb, it is important that they (noun and verb) have a direct relation. Panini has spelled out six karakas (Bharati et al., 1995). Apart from 'karaka' relations, a sentence contains other types of relations as well. For, example, relations such as purpose, reason, genitive etc. These relations are also marked within this scheme.

### **3.2 The Scheme**

The first step in the direction of coming up with a tagging scheme for annotating dependencies at the sentential level was conceived and worked out in 2000 itself. However, it could not be carried forward at the time as other tasks such as POS tagging and chunking etc needed prior attention. In the meanwhile, substantial amount of work has been done in the direction of developing standards for POS tagging and chunking in Indian languages and a tagging scheme for the same (Bharati et al. 2006). It was decided to revisit the AnnCorra Tagset for dependency relations in Jan 2005. Each of the tag was discussed and a revised list was arrived at. The tagset contained around 26 tags.

Based on the tagset developed in 2005, a small set of sentences (about 2000) from Hindi were annotated. During this process it was noted that there were constructions which could not be satisfactorily captured in the existing tagset. Subsequently, the tagset was re-visited and the current tagset was evolved.

The dependency annotation is marked for inter chunk relations. A chunk is taken to be a basic unit for marking the syntactico-semantic relations with the assumption that the intra-chunk dependencies can be obtained automatically by using a rule based system. Also, the verb chunk is more or less a grouping of the verb base form and its tense, aspect and modality (TAM) auxiliaries. The practical aspect of this decision is that it saves the effort in manual annotation.

The tagging scheme for POS and Chunk annotation has been developed through conducting various workshops in which scholars representing nine major languages of India participated. The scheme aimed at coming up with a tagset which would be comprehensive to the extent possible covering issues from all Indian languages and should be simple for the annotators.

Annotation guidelines based on the above scheme are also prepared (Appendix-1). The task of annotating POS and chunk in several Indian languages is already going on under the ILMT project funded by Department of Information Technology (DIT), Ministry of Communication and Information Technology (MCIT), Government of India.

### 3.2.1 Treebank Representation (SSF)

The annotated data is stored in SSF format (Bharati et al., 2007). The SSF is a four column format in which the first column is for address, the second column is for the token, the third column is for the category of the node and the fourth column has other features. Any required linguistic or other information can be annotated in this column using an attribute – value pair. Thus, POS and chunk category of the tokens would be in the third column and the morph, dependency and any other information pertaining to a node would appear in the fourth column.

### 3.2.2 Naming conventions

The naming conventions adopted in the treebank are as follows:

#### A. Naming tokens

Every lexical item and chunk will have a name. The attribute for naming is 'name'. Values for **lexical nodes** would be the lexical item. In case there are more than one occurrences of the same word the value for the name attribute would be the lexical item followed by a numerical. For example, if the token is '*Pala*' (fruit), it would be represented as <name=phala> in the former case. In case '*Pala*' occurs twice in a sentence, the first time its naming feature would be <name=Pala> and the second time it will be named as <name=Pala2>.

#### B. Naming convention for Chunks

The chunks are named as their respective phrase tags(NP/VP/JJP). As in the case of lexical items, the subsequent occurrences of the chunks are also named by appending an iterated number(starting with 2) to the phrase tag.

#### C. Naming the examples in this manual

For ease of access, the examples for various labels and constructions have also been given ids in this document. In PART-1B, the convention is that every example starts with Relation-DS-. Thereafter, the id has the relation label for which the example stands for followed by a number. For example, examples for *karta karaka* would have the following ids – Relation-DS-k1-1, Relation-DS-k1-2 and so on. Similarly, for *karma karaka* examples the ids would be Relation-DS-k2-1, Relation-DS-k2-2 and so on. This allows us a flexibility of adding more examples for each type of relation at a later stage.

In PART-2, the examples are named as [Construction type-DS-examplenumber]. Thus, examples for causative constructions would read as follows : Causative-DS-1, Causative-DS-2 and so on.

### 3.2.3 Relations and Tag labels

The scheme contains about 40 tags which are arrived at considering various types of sentence constructions in Hindi-Urdu. These labels contain relations (a) *karaka* and non-*karaka* dependency relations (b) some underspecified tags of the type *vmod*, *nmod* etc and (c) some tags which indicate relations which are not exactly dependency relations but are required to represent the sentence structures.

As mentioned earlier, the grammatical model captures certain syntactico semantic relations. The tag labels represent various *karaka* and other than *karaka* relations. All *karaka* relations have been labeled starting with a 'k' followed by a numerical. Paninian grammar talks about six *karaka* relations. In this section we describe the *karaka* relations and how they have to be annotated. Although the basic number of *karakas* is six, there are a number of relations which are either finer types of *karakas* (such as *k2p*, *k2g* etc) or are in some way or the other related to a *karaka* (such as *k1s*, *k2s*, *k1u*, *k2u* etc). The labels for dependency relations other than *karakas* start with an 'r'.

There are certain relations which do not fall under 'dependency relation' directly but are required for showing the dependencies indirectly. For example, for representing a labels 'ccof' and 'pof' appear in the tagging scheme to represent 'co-ordination' and 'complex predicates' respectively. The dependency relation type tree in Figure 2 below shows the relations from coarser to finer on a modifier – modified paradigm.

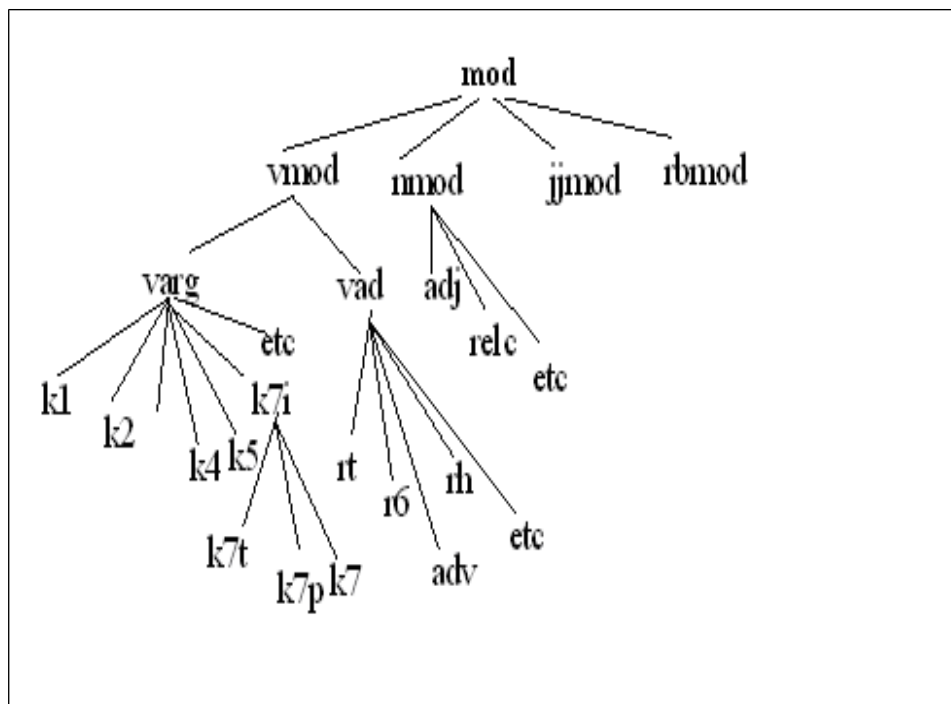


Figure 2

The classification shown in the above tree allows underspecification of certain relations in cases where a finer analysis is not very significant for this level of annotation and is also more difficult for decision making for the annotators. Therefore, the labels such as *k1*, *k2* etc represent a finer level depicted deeper in the tree, whereas, labels such as '*vmod*', '*nmod*' show an underspecified representation of

the relation. More details for this are given under respective labels in Section 4.1 of this document.

In deciding the *karaka* relations of elements in a sentence, the semantics of the verb plays a major role and at the same time syntax helps too. Normally *karta* and *karma* agree with the verb. When *karta* agrees with the verb then it takes *zero* vibhakti otherwise it takes the following vibhaktis: *ne*, *ko*, *se*, *xvArA*. Therefore, a mapping between vibhakti and TAM (tense, aspect and modality) can be quite useful for identifying relations such as *karta* and *karma*.

A default for annotating *karakas* in sentences with more than one verb is that all *karakas* attach to the nearest verb on the right. *kI* has a special default rule for shared *karta* relationship with the verb where it attaches to the finite verb.

### 3.4 Corpora

The Urdu corpus is mainly newspaper texts from *Siyasat*, an Urdu daily published from Hyderabad, Andhra Pradesh, India.

## PART- 1B

The issues related to actual annotation task such as how to mark various relations, how to handle shared arguments, what to do in case of missing arguments are described in this part of the document.

NOTE : Gloss has been provided for the examples given in the document. But often the gloss provides only the relevant lexical information and not all the information which might be there in a word. For example, most often the gender and number information is missing.

### 4.1 How to mark various dependency relations?

We will now take all the relations and the tag labels for the same one by one . A detailed description of every relation and its tag is provided below. The objective of this section is to help the annotators with the actual annotation of various relations in a sentence.

As mentioned above, In this section we have first listed all the relations which have a 'k' label. Then given the labels which have 'r' labels.

#### 4.1.1 **k1** - *Karta* ('doer/agent/subject')

*Karta* is defined as the 'most independent' of all the *karakas* (participants). *Karta* is the one who carries out the action. It is conceptually different from the agent theta role as it does not always have volitionality. It is the locus of the activity implied by the verb root. In other words, the activity resides in or springs forth from the *karta* (Bharati et al., 1995). For example:

1. *Shahid made the basket.*

*Shahid* is *karta* here as he is performing the action of making the basket. In Paninian grammar, every action is a bundle of sub-actions and all the participants (*karakas*) in

an action have a sub-action located in them. Thus every karaka is the 'karta' (doer) of the action located in it. For example;

2.a *Shahid opened the lock with a key*

In the above example, 'Shahid'(karta), 'lock'(karma) and 'key'(instrument) are the three *karakas* of the action of 'opening'. The larger action of opening the lock involves following sub-actions (i) action of Shahid, (ii) action of the lock and (iii) action of the key. Therefore, 'lock' and 'key' are *karta* in examples (2b) and (2c) respectively.

2.b The lock opened

2.c This key opened the lock

The grammar talks of two types of *karta*, (a) primary and (b) secondary. Primary *karta* has volitionality whereas the secondary *karta* does not. Therefore, *kartas* of 2b and 2c above do not have volitionality. In A.B.C. And D. below various conditions under which a '*karta*' occurs in Hindi are explained through examples.

**A.** If the verb denotes an action, the *kl* is the doer of the action. In examples (Relations-k1-1 to 7), ***SAhixa*** is the doer of the action so ***SAhixa*** is the *karta*.

Relation-DS-k1-1 :

***SAhixa*** *bETA hE*  
Shahid sit-perf is  
'Shahid is sitting'

**Syntactic Cues :** Most general or **default syntactic cues** for identifying *karta* in a Hindi-Urdu sentence are:

- (a) *Karta* is normally in nominative case which is realized as 0 in Hindi-Urdu.
- (b) By default *Karta* agrees with the verbal inflection in number, gender and person when the verb is in active voice (list of TAMs attached).

**IMPORTANT NOTE on syntactic cues:** It is important to note that *karta* is not the only *karaka* which may appear with a 0 *vibhakti*. Some other relations may also appear without an explicit case marker. The conditions under which various *karakas* etc occur with a particular '*vibhakti*' may not always be syntactic. Therefore, one may have to use various conditions such as the context, the semantic properties of the word under consideration, semantic properties of the words to which the given word is related etc. In short, the cues provided here are only to help take a decision but are not to be followed fully mechanically.

Some more examples of *karta* with the above syntactic cues are :

Relation-DS-k1-2 :

شاحد کھیر کھاتا ہے  
***SAhixa*** *KIra KAwa hE*  
Shahid rice-pudding eat-hab-sg-m is  
'Shahid eats rice-pudding'

Relation-DS-k1-3 :



نادیا کھیر کھاتی ہے

*nAxiyA KIra KAwi hE*

Nadiya rice-pudding eat-hab-sg-f is

'Nadiya eats rice-pudding'

**B.** However, *karta* in Hindi-Urdu can also occur with case markers other than nominative case (0 vibhakti). The terms case marker, *vibhakti* or postposition are used interchangeably in this document.

Relation-DS-k1-4 :

شاحد نے کھیر کھائی

*SAhixa ne KIra KAyi*

Shahid erg rice-pudding ate

'Shahid ate rice-pudding.'

Relation-DS-k1-5 :

شاحد کو کھیر کھانی پڑی

*SAhixa ko KIra KAnI padZI*

Shahid dative rice-pudding eat+inf+fem had+fem

'Shahid had to eat the rice-pudding'

Relation-DS-k1-6 :

شاحد کو کھیر کھانی چاہیے

*SAhixa ko KIra KAnA cAhiye*

Shahid Dat rice-pudding eat+inf should

'Shahid should eat the rice-pudding'

**Syntactic cues** for identifying a '*karta*' in the above constructions are : If a noun occurs with the postpositions belonging to the list given below and the verb has the corresponding TAM in the list below then the noun would always be a *karta* in Hindi.

**Postposition (Vibhakti)**

(i) *ne*

(ii) *ko*

(iii) *ko*

**TAM**

*yA* (past)

*nA\_padZA*

*nA\_cAhiye*

**C.** In **passive constructions**, normally a *karta* would be absent. However, if it occurs, it will appear either with 'se' as its vibhakti. (Hindi will also take 'xvara' as a Vibhakti)

Relation-DS-k1-7 :

شاحد سے کھیر کھائی گئی

*SAhixa se KIra KAyi gayI*

Shahid by rice-pudding ate Passv

'Rice-pudding was eaten by Shahid.'

**Syntactic cues :** (a) A noun followed by the postposition '*se*' and (b) the verb having a passive TAM (tense, aspect and modality) would be a '*karta*'. A list of passive TAMs

in Hindi-Urdu is provided in Appendix for reference.

**D. Karta with a genitive marker :** Karta in Hindi-Urdu can also occur with a genitive marker. Following are some examples of the same.

Relation-DS-k1-8 :

شاحد کا ماننا ہے کہ کل بارش ہوگی

SAhixa kA mAnanA hE ki kala bAriSa hogI

Shahid of belief is that tomorrow rain will-happen

'Shahid believes that it will rain tomorrow.'

The *karta* with a genitive postposition (kA) occurs only with a few verbs such as 'kaha', 'soca', 'mAna' etc.

**E. Some more examples of 'karta' in Hindi-Urdu sentences**

Relation-DS-k1-9 :

شاحد اچھا ہے

SAhixa acCA hE

Shahid good is

'Shahid is good.'

Relation-DS-k1-10 :

شاحد کو چاند دکھا

SAhixa ko cAzxa xiKA

Shahid dat moon appeared

'Shahid saw the moon.'

In the stative verbs, the state of a person or a thing is mentioned. The person or thing whose state is mentioned will be the *karta*. In example (Relation-DS-k1-8), state of *SAhixa* is mentioned so *SAhixa* becomes the *karta*. The subject of an unaccusative verb would be marked as *karta*. In example (Relation-DS-k1-9), *cAzxa* 'moon' is the *karta* as '*xiKanA*' (to see) is an unaccusative verb in Hindi-Urdu. As mentioned above, *karta* is the doer of the activity denoted by the verb. The activity of '*xeKanA*' (to see) is different from the activity of '*xiKanA*' (to be seen). Therefore, the element from where this activity springs forth would be *karta*.

**F. Clausal *karta* :** A clause can also be *karta*. For example,

Relation-DS-k1-11 :

شاحد کا یہ ماننا سہی نہیں ہے

SAhixa kA yaha mAnanA sahI nahIM hE

Shahid of this belief true not is

'This belief of Shahid is not true.'

In the above example the non-finite clause, '*SAhixa kA yaha mAnanA*' is the *karta* of the verb '*hE*'. The k1 tag in such cases would be annotated on the verb of the clausal *karta*. Therefore, (annotated example is represented in SSF)

(( NP <drel=r6-k1:VGNN> SAhixa NNP kA PSP ))

```
((      NP      <drel=k2:VGNN name=VGNN>      yaha      PRP      ))
((      VGNN      <drel=k1:VGF>      mAnanA      VM      ))
((      JJP      <drel=k1s:VGF>      sahI      JJ      ))
((      VGF      <name=VGF>      nahIM      NEG      hE      VM      ))
```

SSF-1

#### 4.1.2 pk1, jk1, mk1 (causer, causee, mediator-causer)

Causatives in Hindi\_Urdu are realized through a morphological process. An intransitive or a transitive verb changes to a causative verb when affixed by either an '-A' or a '-vA' suffix. In our scheme, both 'causer' and 'causee' are marked. In addition to the causer and causee, there can also be a mediator who is both causee and causer.

##### A. pk1 (prayojaka karta 'causer-1')

Relation-DS-pk1-1 :

ماں نے بچے کو کھانا کھلایا  
*mAz ne bacce ko KAnA KilAyA*  
 mother erg child acc food caused to eat  
 'The mother fed the child.'

Relation-DS-pk1-2 :

نادیا نے آیا سے بچے کو کھانا کھلوايا  
*nAxiyA ne AyA se bacce ko KAnA KilavAyA*  
 Nadiya erg came by child to food caused to eat  
 'Nadiya made the maid to feed the child.'

Relation-DS-pk1-3 :

شاہد نے نادیا سے بھکاری کو پیسے دلوائے  
*SAhixa ne nAxiyA se BiKArI ko pEse xilavAe*  
 Shahid erg Nadiya by beggar acc food caused to give  
 'Shahid made Nadiya give the alms to the beggar'

**Syntactic cues :** Syntactically, 'pk1' will behave like 'karta'. Therefore, all the syntactic cues which are used for 'karta' would apply in the case of a 'prayojak karta' (pk1-causer) as well. The difference between a 'karta' and a 'prayojaka karta' is to be noted from the verb form. '-vA' suffix in the verb is a clear indicator of it being a causative.

##### B. jk1 (prayojya karta 'causee')

The causee in a causative construction is annotated as **jk1**. All the tags capture the information of agentive participation in various nouns.

Relation-DS-jk1-1 :

ماں نے بچے کو کھانا کھلوايا  
*mAz ne bacce ko KAnA KilavAyA*  
 mother erg child acc food caused to feed  
 'The mother made (someone) feed the child.'

Relation-DS-jk1-2 :

شاحد نے نادیا سے ٹکٹ خریدوائے

*SAhixa ne nAxiyA xvArA/se tikata KArixavAye*

Shahid erg Nadiya by ticket caused to buy

'Shahid made Nadiya to buy tickets.'

Relation-DS-jk1-3 :

شاحد نے نادیا سے آمر کو ٹکٹ دلوائے

*SAhixa ne nAxiyA xvArA/se Amira ko tikata xilavAye*

Shahid erg Nadiya by Amir Dat ticket caused to give

'Shahid made Nadiya to buy tickets for Amir.'

**Syntactic cues :** Syntactically, a causee would have either a 'ko' vibhakti or a 'se' vibhakti. The choice of 'ko' or 'se' would depend on the type of verb. Therefore, there is no definite syntactic cue. In this case also, it is the verb form and its semantics which are the determining factors for identifying this relation.

**C. mk1** (madhyastha karta 'mediator causer')

Causative constructions have at least one causer and one causee. However, more than one causers can also occur in a sentence. The second causer in such cases is a causee-causer. The causee-causers can be more than one in a sentence. See the examples below :

Relation-DS-mk1-1 :

ماں نے آیا سے بچے کو کھانا کھلوا یا

*mAz ne AyA se bacce ko KAnA KilavAyA*

mother erg Ayah by child acc food made to eat

'The mother made the Ayah to make the child eat the meal'

Relation-DS-mk1-2 : *SAhixa ne Saif ke zariye nAxiyA se BiKArl ko KerAwa xilavAyA*

Shahid erg Saif by Nadiya by beggar Dat food caused to give

'Shahid made Saif to make Nadiya give the alms to the beggar'

Relation-DS-mk1-3 : *nAxiyA ne mariyama ke zariye AyA se bacce ko KAnA KilavAyA*

Nadiya erg Maryam by maid by child acc food caused to feed

'Nadiya made Maryam to make the maid feed the child.'

**Syntactic cues :** The vibhakti for a 'mk1' would either be 'xvArA' or 'se'. When more than one mk1 occurs in a sentence, then the first one would have 'ke\_zariye' vibhakti and the second one would have 'se' vibhakti.

Therefore, the causer – causee relation is derived from the verb morphology rather than other clear syntactic cues.

#### 4.1.3 k1s (vidheya karta - karta samanadhikarana 'noun complement of karta')

Noun complements of *karta* are marked as 'k1s'. The term *samanadhikarana* indicates 'having the same locus'. Therefore, *karta samanadhikarana* indicates having the same locus as *karta*.

Relation-DS-k1s-1 :

شاحد عقل مند ہے  
SAhixa aqalamanxa hE  
Shahid intelligent is  
'Shahid is intelligent.'

Relation-DS-k1s-2 :

سپاہی بڑا بہادر ہے  
sipAhi baRA bahAxura hai  
soldier very brave is  
'The soldier is very brave.'

#### 4.1.4 k2 (karma 'object/patient')

The element which is the object/patient of the verb is marked as *karma*. *Karma* is the locus of the result implied by the verb root.

##### A. karma in active voice sentences:

Giving below are some examples of the occurrence of *karma* in active voice sentences:

Relation-DS-k2-1 :

شاحد روز ایک سیب کھاتا ہے  
SAhixa rojZa eka seba KAwa hE  
Shahid everyday one apple eat-hab pres  
'Shahid eats an apple everyday'

Relation-DS-k2-2 :

شاحد نے کل ایک فلم دیکھی  
SAhixa ne kala eka Pzilma xeKI  
Shahid erg yesterday one film saw  
'Shahid watched a movie yesterday'

Relation-DS-k2-3 :

شاحد نے بازار میں حامد کو دیکھا  
SAhixa ne bAjZAra meM hAmixa ko xeKA  
Shahid erg market in Hamid acc saw  
'Shahid saw Hamid in the market'

**Syntactic cues :** *Karma* occurs either with a zero vibhakti or a 'ko' vibhakti. Often, in Hindi-Urdu, both *karta* and *karma* would occur without a postposition/vibhakti (zero

vibhakti). When both *karta* and *karma* occur with a zero vibhakti in a sentence, then the two nouns are of different gender then the noun which **does not** agree with the verb would be *karma* (see example Relation-DS-k2-1 above). If the *karta* in a sentence is marked by a postposition, then the noun which agrees with the verb would be *karma* (Relation-DS-k2-2). *Karma* can also occur with a 'ko' vibhakti. *Karma* would be marked by a 'ko' vibhakti when it is a human noun (Relation-DS-k2-3). Sometimes, *karma* is marked by a 'ko' vibhakti to indicate definiteness.

**B. In passive constructions**, the noun which agrees with the verb is the *karma*.

Relation-DS-k2-4 :

دیوالی پر خوب مٹھائی کھائی گئی  
*xivAll para KUba miTAI KAyI gayI*  
 Diwali on lots of sweets eat-Passv  
 'Lots of sweets were eaten on Diwali'

Relation-DS-k2-5 :

دیوالی پر خوب پٹا خنے چھوڑے گئے  
*xivAll para KUba pataKe CodZe gaye*  
 Diwali on lots of crackers leave go-Passv  
 'Lots of crackers were burst on Diwali'

**Syntactic cues :** If the verb in a sentence occurs with a passive TAM then the noun which agrees with the verb is the *karma*

**C. Vakya-karma** (Sentential object – 'complement clauses')

Finite clauses occur as sentential object the verb of the subordinate clause is attached to the verb of the main clause and the arc is tagged as 'k2'. For example,

Relation-DS-k2-6 :

شاحد نے بتایا کہ باہر پانی برس رہا ہے  
*SAhixa ne bawAyA ki bAhara pAnI barasa rahA hE*  
 Shahid erg told that outside water raining prog pres  
 'Shahid told that it was raining outside'

Relation-DS-k2-7 :

اس نے کہا کہ شاحد کل نہیں آئے گا  
*usane kahA ki SAhixa kala nahIM AyegA*  
 he-erg told that Shahid tomorrow not will-come  
 'He told that Shahid will not come tomorrow.'

#### 4.1.5 k2p (Goal, Destination)

The destination or goal is also taken as a *karma* in this framework. However, it is marked as *k2p* in the treebank. *k2p* is a subtype of *karma* (k2). The goal or destination where the action of motion ends is a *k2p*. These are mostly the objects of motion verbs. They also occur with other types of verbs. The syntactic behavior of *k2p* is slightly different from other *k2*. That is why a separate tag has been kept for them.

Unlike other *karma*, the goal/destination *karma* do not agree with the verb under similar syntactic context (see example Relation-DS-k2p-2 below).

Relation-DS-k2p-1 :

شہاد گھر گیا  
*SAhixa Gara gayA*  
 Shahid home went  
 'Shahid went home.'

Relation-DS-k2p-2 :

وہ سب کو اپنے گھر بلاتا ہے  
*vaha saba ko apane Gara bulAwA hE*  
 he all acc his home invite be-Pres  
 'He invites everybody to his home.'

Relation-DS-k2p-3 :

شہاد کو دلی جانا پڑا  
*SAhixa ko xillI jAnA padZA*  
 Shahid acc Delhi go lie  
 'He had to go to Delhi'

Relation-DS-k2p-3 b : \* *SAhixa ko xillI jAnI padZI*

'xillI' is a feminine noun in Hindi-Urdu. However, an agreement between 'xillI' and the verb 'jAnA padZA' in example Relation-DS-k2p-3 b above is ungrammatical. This is why, though a destination is also a *karma*, it is treated as a special case.

#### 4.1.6 k2g (secondary karma)

It is possible to have more than one 'karma' of the same verb in a sentence. For example:

Relation-DS-k2g-1 :

وہ لوگ گاندھی کو باپو بھی کہتے ہیں  
*ve loga gAMXIjI ko bApU BI kahawe hEM*  
 those people Gandhi+hon acc Bapu also say+hab be-Pres  
 'They also call Gandhiji as Bapu.'

Verbs such as 'kahanA' can have two *karma*. In sentence Relation-DS-k2g-1 above, 'kahate hEM' has two *karmas* - 'gAMXIjI' and 'bApU'.

#### 4.1.7 k2s (karma samanadhikarana 'object complement')

The object complement is called as *karma samanadhikarana* and the tag used for it is 'k2s'.

Relation-DS-k2s-1 :

وہ گاندھی کو باپو بھی مانتے ہیں  
*ve gAMXIjI ko bApU BI manawe hEM*

they Gandhiji acc father also believe+hab be-Pres  
'They consider Gandhiji as a father.'

Relation-DS-k2s-2 :

شاحد نادیا کو عقل مند سمجھتا ہے

*SAhixa nAxiyA ko aqalamanxa samaJawA hE*  
Shahid Nadiya acc intelligent consider-Impf be-Pres  
'Shahid considers Nadiya to be intelligent.'

Notice that both *kahana* 'to say' and *mAnana* 'to believe' seem to have two karmas, but only 'kahana' can be treated as taking two 'karma'. This is because in (Relation-DS-k2g-1), *bApU* 'bapu' is a word or substance, whereas in (Relation-DS-k2s-1), *bApU* 'bapu' is a property that resides in *gAMXljl*. That is why in Relation-DS-k2s-1 *bApU* is the object of a ditransitive verb and in Relation-DS-k2s-1 'bApU' is the k2s.

#### 4.1.8 k3 (karana 'instrument')

*karana* karaka denotes the instrument of an action expressed by a verb root. The activity of *karana* helps in achieving the activity of the main action. For example,

Relation-DS-k3-1 :

شاحد نے چاقو سے سیب کاٹا

*SAhixa ne cAkU se seba kAtA*  
Shahid erg knife inst apple cut  
'Shahid cut the apple with a knife.'

The element 'with a knife' in the above sentence is *karana* as with the help of the knife, the result, i.e. the 'pieces of the apple', is achieved. Some more examples of sentences having *karana karaka* are given below.

Relation-DS-k3-2 :

شاحد نے چم سے کھیر کھائی

*SAhixa ne cammaca se KIra KAyI*  
Shahid erg spoon with rice-pudding ate  
'Shahid ate the rice-pudding with a spoon.'

Relation-DS-k3-3 :

نادیا نے پانی سے گھڑے کو بھرا

*nAxiyA ne pAnI se GadZe ko BarA*  
Nadiya erg water with clay-pot acc filled  
'Nadiya filled the clay-pot with water.'

Relation-DS-k3-4 :

شاحد نے رسی کے ذریعے سامان باندھا

*SAhixa ne rassI ke zariye samAna banXA*  
Shahid erg rope with luggage tied  
'Shahid tied the luggage with rope.'

Any element/noun which is instrumental in achieving the result would be marked as



'k3' for *karana*. The noun need not necessarily denote a physical object which is an instrument. For example, the noun 'pAnI' (water) in the sentence Relation-DS-k3-3, is instrumental in achieving the action of 'BaranA' (to fill). Thus, 'pAnI' (water) would be marked as 'k3' (*karana*).

**Syntactic cues :** *karana* karaka always takes a *se* or *a ke zariye* vibhakti in Urdu.

#### 4.1.9 k4 (sampradana 'recipient')

*Sampradana* karaka is the recipient/beneficiary of an action. In other words, the person/object for whom the *karma* is intended for is *sampradana*.

Relation-DS-k4-1 :

شاحد نے نادیا کو کھیر دی  
*SAhixa ne nAxiyA ko KIra xI*  
 Shahid erg Nadiya dat rice-pudding gave  
 'Shahid gave rice-pudding to Nadiya.'

Relation-DS-k4-2 :

شاحد نے نادیا کو کہانی سنائی  
*SAhixa ne nAxiyA ko kahAnI sunAyI*  
 Shahid erg Nadiya dat story told  
 'Shahid narrated a story to Nadiya.'

Relation-DS-k4-3 :

میں نے احمد کو کاجل لگایا  
*mEne ahamaxa ko kAjala lagAyA*  
 I erg Ahmed dat mascara put  
 'I put mascara on Ahmad.'

The final destination of the action *xI* 'gave' in Relation-DS-k4-1 above is *nAxiyA* 'Nadiya' which is marked with *ko*. Similarly the final destination of the action *sunAyI* 'told' in Relation-DS-k4-2 is *nAxiyA* 'Nadiya' which is again marked with *ko*.

**Syntactic cue :** *sampradana* karaka normally takes a *ko* or *ke* vibhakti in Urdu.

**B.** Certain cases where *sampradana* does not take a 'ko' or 'ke' postposition

Verbs such as 'kahanA' take a 'se' vibhakti for K4.

Relation-DS-k4-3 :

شاحد نے حامد سے یہ کہا  
*SAhixa ne hAmixa se yaha kahA*  
 Shahid erg Hamid to this said  
 'Shahid said this to Hamid.'

It appears that some communication verbs take 'se' vibhakti for k4 but not all. Therefore, k4 of verbs such as 'bawAnA', 'sunAnA' does not take a 'se' vibhakti. It takes a 'ko' vibhakti in these cases also.

Relation-DS-k4-4 :

شاحد نے حامد کو یہ بات بتائی  
*SAhixa ne hAmixa ko yaha bAwa bawAyI*  
Shahid erg Hamid to this matter told  
'Shahid told this (matter) to Hamid.'

#### 4.1.10 k4a (anubhava karta 'Experiencer')

In perception verbs such as *seems*, *appear* there is a perceiver/experiencer. In the example Relation-DS-k4a-1 below, *SAhixa* is *k1*, *aqalamanxa* is *k1s* and *muJako* 'I-Dat' is *k4a* (perceiver). Here *muJako* 'I-Dat' is a passive agent i.e. experiencer who is not making any effort but just receiving or perceiving the activity carried out by another agent is identified as *anubhava karta* and is marked as *k4a*. The term *anubhava karta* does not occur in Sanskrit grammatical literature. This has been introduced here for Hindi based on the observations of Hindi syntax. Also, since the passive participation of perceiving is that of a recipient, it has been placed under *sampradana* here. The *anubhava karta* can be equated with dative subject.

Relation-DS-k4a-1 :

مجھ کو شاحد عقل مند لگتا ہے  
*muJako SAhixa aqalamanxa lagawA hE*  
I-Dat Shahid intelligent seems be-Pres  
'Shahid seems intelligent to me.'

**Syntactic cues :** *anubhava karta* always takes a *ko* or *ke* vibhakti. Argument of unaccusative verbs having a 'ko' vibhakti would also be marked as *anubhava karta* (Example Relation-DS-k4a-2 below). Verbs such as *laganA* 'to seem' and *xiKAnA* 'to appear' take passive agents and would be marked as 'k4a'. On the other hand, verbs such as *mAnanA* 'to believe' and *xeKAnA* 'to see' take active agents and would be marked as 'k1'. See the following examples:

Relation-DS-k1-12 :

وہ مانتا ہے کہ شاحد عقل مند ہے  
*vaha mAnawA hE ki SAhixa aqalamanxa hE*  
he believe+hab be-Pres that Shahid intelligent be-Pres  
'He believes that Shahid is intelligent.'

Relation-DS-k4a-2 :

مجھ کو چاند دکھا  
*muJako cAzxa xiKA*  
I-Dat moon appeared  
'I saw the moon.'

Relation-DS-k1-13 :

میں نے چاند دیکھا  
*mEne cAzxa xeKA*  
I-Erg moon saw  
'I saw the moon.'

In examples (Relation-DS-k1-12 and 13), *vaha* 'he' and *mEne* 'I-erg' respectively are *k1* as they are active agents. On the other hand, in examples Relation-DS-k4a-1 and 2, *muJako* 'I-Dat' is *k4a* as in both the examples it appears as a passive agent (experiencer). Some more examples of *anubhava karta* are:

Relation-DS-k4a-3 :

شاحد کو کتاب ملی  
*SAhixa ko kiwAba mill*  
 Shahid Dat book got  
 'Shahid found a book.'

Relation-DS-k4a-4 :

شاحد کو بھوک لگی  
*SAhixa ko BUka lagI*  
 Shahid Dat hungry felt  
 'Shahid felt hungry.'

Relation-DS-k4a-5 :

مجھے / مجھ کو اچھا لگا  
*muJe/muJako acCA lagA*  
 I-Dat good felt  
 'I felt good.'

Relation-DS-k4a-6 :

مجھے / مجھ کو لڈو اچھے لگتے ہیں  
*muJe/muJako laddU acCe lagawe hEM*  
 I-Dat sweet good feel-hab be-Pres  
 'I like sweets.'

Relation-DS-k4a-7:

احمد کو تھپڑ پڑا  
*ahamaxa ko TappaDa paDA*  
 Ahmad -Dat slap got  
 'Ahmad was slapped.'

#### 4.1.11 k5 (*apadana* 'source')

*apadana karaka* indicates the source of the activity, i.e. the point of departure. A noun denoting the point of separation for a verb expressing an activity which involves movement away from is *apadana*. In other words, the participant which remains stationary when the separation takes place is marked *k5*.

Relation-DS-k5-1 :

شاحد نے چمچ کے ذریعے کٹوری سے کھیر کھائی  
*SAhixa ne cammaca ke zariye katorI se KIra KAyI*  
 Shahid erg spoon with bowl from rice-pudding ate  
 'Shahid ate the rice-pudding from a bowl with a spoon.'

Relation-DS-k5-2 :

چور پولس سے بھاگتا ہے  
cora pulisa se BAgaWA hE  
thief police from run-away-hab pres  
'The thief runs away from the police.'

**Syntactic cues :** *apadana karaka* always takes a *se* vibhakti in Hindi. However, since 'se' postposition in Hindi is functionally overloaded, it is not a very reliable cue for identifying a *karaka*. Therefore, one has to look for additional cues in cases where 'se' is a vibhakti. The other cue in case of *apadana karaka* would be the verb semantics. If the verb denotes some motion, then the point of departure would be marked with 'se' and that would be *apadana karaka*.

**B.** Emotional verbs such as *gussa honA* 'to be angry', *KuSa honA* 'to be happy' also take an *apadana karaka*. The entity which triggers these emotions is annotated as k5

Relation-DS-k5-3 :

شاحد نادیہ سے غصہ ہے  
SAhixa nAxiyA se gussa hE  
Shahid Nadiya from angry is  
'Shahid is angry with Nadiya'

The example Relation-DS-k5-3 shows a case where there is no explicit point of separation from the noun 'nAxiyA' (Nadiya). However, it will still be marked as 'k5' since it expresses the source of anger. At an abstract level, the anger is triggered from Nadiya. Thus, 'nAxiyA' (Nadiya) would be the point of departure for the emotion of anger triggered in 'SAhixa' (Shahid) and will be marked as 'k5'.

**C.** Verbs such as *pUCanA* 'to ask' also take a k5. The entity from which the information has to be elicited is marked as k5 as it functions as the source.

Relation-DS-k5-4 :

میں نے اس سے ایک سوال پوچھا  
mEMne usase eka savAla pUCA  
I-erg him-abl one question asked  
'I asked him a question.'

#### 4.1.12 k5prk (prakruti apadana 'source material' in verbs denoting change of state)

Example such as the following pose an interesting problem for appropriate *karaka* assignment.

Relation-DS-k5prk-1 :

جوتے چمڑے سے بنتے ہیں  
jUwe camade se banawe hEM  
shoes leather from make-hab be-pres-pl  
'The shoes are made of leather.'

The issue here is whether 'camade' (leather) in the above example is *karana karaka* or

*apadana*. Both these *karakas* in Hindi take a 'se' postposition. Therefore, how do we decide what role 'camade' (leather) is playing in the action of 'banate' (make). An instrument participates in an action as a mediator for accomplishing the result of the action and is not itself affected by it, i.e., it does not undergo a change. However, 'camade' as a participant in the action of 'banate' (make) undergoes a change and also has a relation with the finished product. Change of state verbs such as 'make' require at least two participants 'a raw material' ('leather' in this case) with the aid of which a finished product ('shoes' in this case) is made. Hence, it is a relation which involves a kind of separation – separation from the larger raw material from which a product is made. The karaka relation will then be a special case of apadaan i.e k5. This is because there is a conceptual separation point from the original raw material 'camade' (leather) to the finished product 'jUte' (shoes). The two states in this change of state action are referred to as *prakriti* 'natural' and *vikruti* 'change'. Therefore the tag for this type of *apadana* is named as 'k5prk'.

#### 4.1.13 k7t (kAlAdhikarana 'location in time')

Adhikaran karaka is the locus of karta or karma. It is what supports, in space or time, the karta or the karma. The participant denoting the time of action is marked as 'k7t'. *adhikarana* can be of time or space. It is not mandatory of *adhikarana* to always take a vibhakti. Therefore, even k7t may occur with or without a vibhakti. For instance, in example Relation-DS-k7t-2 and 3 there are no vibhaktis, whereas Relation-DS-k7t-4 and 5 take a *meM*.

Relation-DS-k7t-2 :

کل پانی برساتا تھا

*kala pAnI barasA WA*  
yesterday water rained be-Past  
'It rained yesterday.'

Relation-DS-k7t-3 :

شہید پہلے آیا

*SAhixa pahale AyA*  
Shahid first came  
'Shahid came first.'

Relation-DS-k7t-4 :

اس زمانے میں مہنگائی کم تھی

*usa jZamAne meM mahazgAI kama WI*  
that period in expensive-ness less be-Past  
'The cost of living was less those days'

Relation-DS-k7t-5 :

بچپن میں وہ بہت شرارتی تھا

*bacapana meM vaha bahuwa SarArawi WA*  
childhood in he very naughty be-Past  
'He was very naughty in his childhood.'

**Syntactic cue :** As mentioned above, 'k7t' is often marked by a 'meM' vibhakti. Some time expressions (such as 'subaha' – morning, 'pahale' – before/first, 'kala' –

yesterday/today, 'mahIne' – month etc) when participating in an *adhikarana* role do not take any vibhakti. However, there are some specific cases where 'k7t' has other vibhaktis as well. For example,

Relation-DS-k7t-6 :

تم میرے گھر شام کو آنا  
*wuma mere Gara SAma ko Ana*  
 you my home evening acc come  
 'You come to my place in the evening.'

Relation-DS-k7t-7 :

شاحد اپنا کام وقت پر کرتا ہے  
*SAhixa apanA kAma vaqwa para karawA hE*  
 Shahid own work time on do-hab be-pres  
 'Shahid does his work on time'

#### 4.1.14 k7p (deshadhikarana 'location in space')

The participant denoting the location of *karta* or *karma* at the time of action is called as *deshadhikarana*. It will be marked as 'k7p'. Some examples of 'k7p' are given below.

Relation-DS-k7p-1 :

میز پر کتاب ہے  
*mejZa para kiwAba hE*  
 table on book is  
 'The book is on the table.'

Relation-DS-k7p-2 :

ہوا میں ٹھنڈک ہے  
*havA meM TaMdaka hE*  
 air in chill is  
 'The air is very chill.'

Relation-DS-k7p-3 :

شاحد وہاں کھڑا ہے جہاں نادیہ کھڑا ہے  
*SAhixa vahAz KadZA hE jahAz nAxiyA KadZA hE*  
 Shahid there standing is where Nadiya standing is  
 'Shahid is standing there where Nadiya is standing.'

Relation-DS-k7p-4:

احمد وادی کشمیر کو آیا  
*ahamaxa vAxi-e-kaSamIra ko AyA*  
 Ahmad valley of Kashmir came to  
 'Ahmad came to the valley of Kashmir.'

**Syntactic cues :** Like location of time(k7t), some locations of place carry explicit vibhaktis (case markers) and some don't. When a location of place does take an explicit vibhakti then most of the postposition would be **ko 'to'**, *meM* 'in' or *para* 'on'. In example Relation-DS-k7p-3 'k7p' has no vibhakti. The tag k7p refers to a location of place which is an **actual physical place and not a metaphorical or abstract place**.

#### 4.1.15 k7 (vishayadhikarana 'location elsewhere')

Another kind of *adhikarana* is *vishayadhikarana* which can be roughly translated as 'location in a topic'. For example

Relation-DS-k7-1 :

وہ سیاست پر گفتگو کر رہے ہیں  
*ve siyAsawa para guPawagU kara rahe haiN*  
 they poilitics on discussion do prog be  
 'They are discussing politics.'

However, the term 'topic' can be misleading as it is not restricted to the 'topic' of discourse alone. It is in fact a location other than time and place. Some more examples of *vishayadhikarana* are :

Relation-DS-k7-2 :

حامد نے آزادی کی جنگ میں حصہ لیا  
*hAmixa ne svawanwrawA saMgShahida meM hissA liyA*  
*Hamid erg independence movement in part took*  
 'Hamid took part in the independence movement.'

Relation-DS-k7-3 :

تاریخ میں اس کی مثال نہیں ملتی  
*waRIkha meN isa ki misAla nahiN milwI*  
 history in this of example not found  
 'An example of this is not found in history.'

Relation-DS-k7-4 :

میرے من میں غصہ ہے  
*mere mana meM gussA hE*  
 my mind in anger is  
 'I am angry'

Relation-DS-k7-5 :

میرا من امریکا میں ہے  
*merA mana amarIkA meM hE*  
 my mind America in is  
 'I am mentally in America.'

In the example (4) above '*mana*' is not a concrete physical place, therefore, it will be marked as k7. In the example (5), '*amerikA*' is an actual physical place, but this will also be NOT marked as k7p. Instead, it will be marked as k7. The reason for marking

it as k7 is that though America is an actual physical place, but the entity (*mana* in this case) which is in America is not. So, for a participant to be marked as k7p there has to be an actual physical contact, i.e., the located and the location have to be concrete objects. If they are not, then the location would be marked as k7.

**Syntactic cue :** Like other types of *adhikarana*, *vishayAdhikarana* also takes 'meM' and 'para' postpositions as its case markers.

#### 4.1.16 k\*u (sAdrishya 'similarity')

The tag to mark similarity is 'k\*u'. This can be used for marking both similarity and comparison. The tag is marked on the comparand in a comparative construction. Since the compared entity can compare with any *karaka*, the tag includes a star. '\*' in the tag name is a variable for whichever *karaka* is the comparee of the comparand. Therefore, while marking the comparand (the compared entity), the \* would be replaced by the appropriate *karaka* label. For example,

Relation-DS-k\*u-1 :

رادھا میرا جیسی خوبصورت ہے  
*SAziyA mariyama jEsI khUbasUrawa hE*  
 Shaziya Maryam like beautiful is  
 'Shaziya is beautiful like Maryam.'

In the above example, '*SAziyA*' is the karta of the verb '*hE*'. '*mariyama*' is the comparand (entity with which '*SAziyA*', the karta, is being compared) and '*SAziyA*' is the comparee (entity which is being compared). Therefore, '*mariyama*' in the above example will be annotated as 'k1u'. Some more examples are given below:

Relation-DS-k\*u-2 :

نادیا کو مریم شازیہ جیسی خوبصورت لگی  
*nAxiyA ko mariyama SAziyA jEsI khUbasUrawa lagI*  
 Nadiya Dat Maryam Shaziya like beautiful appeared  
 'To Nadiya Maryam appeared as beautiful as Shaziya.'

Relation-DS-k\*u-3 :

نادیا مریم کو شازیہ جیسی خوبصورت مانتی ہے  
*nAxiyA mariyama ko SAziyA jEsI khUbasUrawa mAnatI hE*  
 Nadiya Maryam acc Shaziya like beautiful consider pres  
 'Nadiya considers Maryam as beautiful as Shaziya.'

Relation-DS-k\*u-4 :

شازیہ مریم کے مقابلے زیادہ خوبصورت ہے  
*SAziyA mariyama ke muqAbale zyAxA khUbasUrawa hE*  
 Shaziya Maryam of comparison more beautiful is  
 'Shaziya is more beautiful in comparison to Maryam.'

Relation-DS-k\*u-5:

احمد بیل کی مانند محنت کرتا ہوا دیکھا گیا  
*ahamaxa bEla ki mAninxA mehanawa karwA xeKA gAyA*



Ahmad bull like hard-work do seen  
'Ahmad was seen working hard like a bull.'

Similarly, in the example Relation-DS-k\*u-2, 'Maryam' is the comparee and 'Shaziya' comparand. Therefore, 'Shaziya' would be marked as 'k1u'. However, in example Relation-DS-k\*u-3, 'Maryam', the comparee is 'k2', thus 'Shaziya', the comparand will be annotated as 'k2u'.

**Syntactic cue :** In the comparative constructions the comparand will take either 'jEsA' or 'se' postposition.

#### 4.1.17a r6 (shashthi 'genitive/possessive')

The genitive/possessive relation which holds between two nouns has to be marked as 'r6'. For example,

Relation-DS-r6-1 :

کتاب کی قیمت  
*kiwAba kI kImawa*  
book of price  
'Price of the book.'

**Syntactic cues :** This is one of the easy to identify relations in Urdu when it occurs with a Vibhakti. It has a relatively reliable syntactic cue. It mostly occurs with a 'kA' postposition. A reliable cue for its identification is that the postposition 'kA' agrees with the noun it modifies in number and gender. Thus, in example Relation-r6-1 above 'kA' has masculine gender and singular number which agrees with the following noun (its modified) ". In Relation-r6-1 , the postposition 'kA' agrees with 'kImawa' , a feminine noun in Hindi-Urdu.

#### 4.1.17b r6 (ezafe)

Definition: Ezafe is a marker especially a sound that is represented by a character that facilitates the link or the relation between two words of the same constituents.

The inter linked words may be named the ezafe (-e) construction. Note that ezafe is not always, orthographically rendered in Urdu .The ezafe character that creates 'ye/-e-' sound is very often, dropped orthographically. It is best verified if spoken.

The Urdu ezafe construction, a borrowing from Persian and Arabic, expresses the dependency between a noun-noun pair,noun-adjective pair,adjective-noun pair .

The use of ezafe is not restricted to high literary language,it has become a part and parcel in the communication process of contemporary generation.

The terms, Dard-e-sar, dard-e- Disco, wakht-e-maut, nigaah-e-karam are liberally used.

Given below is an example of the ezafe construction in Urdu

Relation-DS-r6-2 :

صدر ریاست دلی پہنچے

*saxara-e riyAsawa dilli pahunce*

president of country delhi arrive-p

'The President of the country arrived in Delhi.'

Contrary to the usual SOV and therefore, head-final nature of Urdu NP constructions, the ezafe construction is head-initial in terms of word order, i.e. the syntactic head is to the left. Contrast this with the following example which is semantically similar to the one above, and uses the common genitive *kA* to the same purpose.

Relation-DS-r6-3:

ریاست کے صدر دلی پہنچے

*riyAsawa ke saxara dilli pahunche*

country of president delhi arrive-p

'The president of the country arrived in Delhi.'

The example in Relation-DS-r6-2 uses the ezafe construction for its NP, while the one in Relation-DS-r6-3 uses the genitive *kA* for its NP. The word order inside the NP as mentioned above, differs in both cases.

It is not clear as to what kind of noun-noun pairings does the ezafe construction license. Can it be used freely and productively to pair any two nouns that might come to bear an r6 relation, or is there a certain limitation to this licensing ? We intend to explore this in future. For now, we rely on the annotators familiarity with the constructions (that are used fairly regularly in texts), to mark them as r6. Agreement with the verb often acts as a syntactic clue for the occurrence of such constructions. A few more examples of the construction are given below:

Relation-DS-r6-4 :

ہم نے ملکہ غزل بیگم اختر کو سنا

*hamne mallikA-e ghazala begama akhwara ko suna*

we -erg queen of ghazal begum akhtar listen

'We listened to Begum Akhtar, the queen of ghazal.'

Relation-DS-r6-5:

شام اودھ کی اپنی ہی بات ہے

*Sama-e avaXa ki apni hi bAwa hE*

evening of avadh its charm is

'The evening of Avadh has its own charm.'

The character or marker that links up a pair of words in order to show some relation is ezafat. The constituent of such pairing can be considered ezafat construction. Ezafa constructions are found to have the following various forms of pairing up.

1. Noun -Noun pairing

2. Adjective-Noun

### 3. Noun -adjective

### 4. Spatial or temporal Noun- Noun

The dependency structure for each of the ezafe constructs are displayed below:

#### 1. Noun-noun

For example:

بی جے پی کے ایک رکن اسمبلی نے لوک سبھا انتخابات میں کامیابی حاصل کی۔

B.J.P. Ke ek rukn-e-assembly ne lok saBA se kAmyAbi hAsila kI

B.J.P. -gen one member of assembly -erg lok sabha elections in victory gain past  
a member of assembly of BJP has gained victory in the elections.

رکن اسمبلی rukn assembly ezafat -e- is dropped orthographically,  
hence ,the construction is divided conveniently where head initial  
property of ezafa construction can be captured.

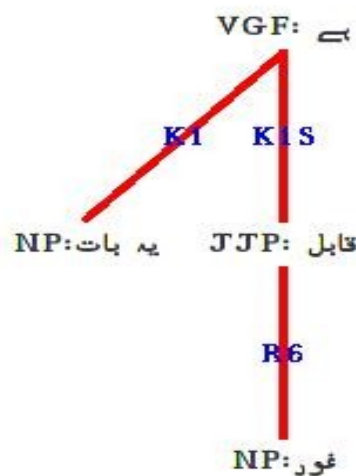
[(BJP)NN( ke)PSP]NP  
[(ek)NN]NP2 {NMOD:NP3}  
[(rukn) NN]NP3 {k1:VGF}  
[(assembly)NN]NP4 {r6:NP2}  
[(ne)PSP]fragp {fragof;NP3}

#### 2. Adjective-Noun

یہ بات قابل غور ہے

ye bAw khAbil-e- ghaur hE  
this matter worth notice pres  
this matter is worth notice.

[(ye)DEM(bAw)]NP[(Qabil)]JJJJP [(ghaur)NN]NP [(hE)VGF



### 3. Noun-adjective

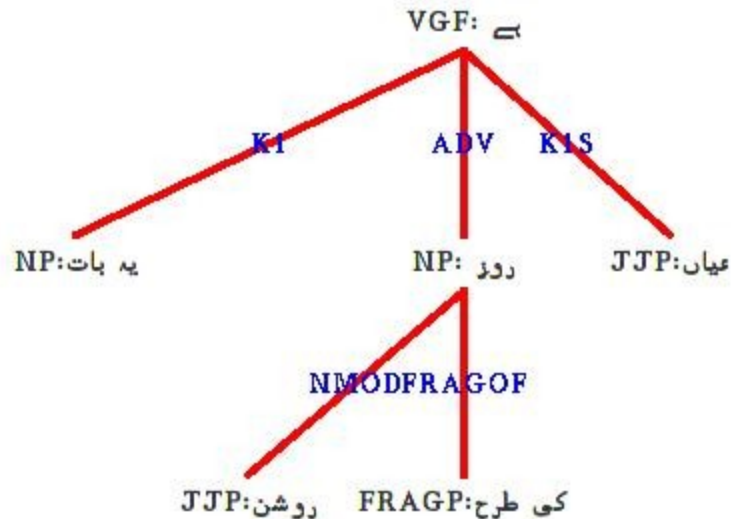
یہ بات روز روشن کی طرح عیاں ہے۔

ye bAw roz(-e-)roSana kii taraha ayAn hE

this matter day illuminate-past participle of kind obvious is  
this matter is obvious like the illuminated Day.

[(ye)DEM(bAw)NN]NP {k1:VGF} [(roz)NN]NP2 {adv:VGF}

[(roSana)JJ]JJP {nmod:VGF} [(kii)[JJP {k1s:VGF} [hE(VGF)



### 4. NST-Noun

Spatial and temporal Noun

بیرون ملک مقیم افراد نے بھی مبارکباد پیش کی۔

baerUn-e-mulk muqIm afrAd ne Bi mubArakbAd peS kI

outside nation resident people -erg too congratulation present-past  
people residing outside India too congratulated.

The above discussed ezafa construction can be chunked as a single NP since the constituent is made of our NST baerun that is outside and mulk that is nation Noun. Since ezafa is usually dropped orthographically it need not to be marked. Such ezafa constructions do not require an alternative method of marking dependency structure.

#### 4.1.18 r6-k1, r6-k2 (karta or karma of a conjunct verb (complex predicate))

Indian languages make extensive use of conjunct verbs. A conjunct verb is composed of a noun or an adjective followed by a verbalizer. Some times the argument (*karta or karma*) occur in a genitive case. Whenever the argument of a conjunct verb is in genitive case it will have a dependency relation with the noun of the conjunct verb. This is because the argument in the genitive case agrees with the noun of the conjunct verb and not with the verb. The noun of the conjunct verb agrees with the verb. In the

example Relation-DS-r6-k1-1 below, *maMxira kA* ‘temple of’ will be marked as *r6-k1* with *uxGAtana* ‘inauguration’. *maMxira* has *r6* relation with the noun of conjunct verb and in the sentence, *maMxira* has karaka relation *k1 of the conjunct verb* ‘*uxGAtana karanA*’. In example Relation-DS-r6-k2-1, *maMxira kA* ‘temple of’ will be marked as *r6-k2* with *uxGAtana* ‘inauguration’. *maMxira* has *r6* relation with the *uxGAtana* ‘inauguration’ which is the noun of conjunct verb and in the sentence, *maMxira* has karaka relation of *k2*.

Relation-DS-r6-k1-1 :

اس چوراہے سے یومیہ 20 ہزار سے زائد گاڑیوں کی آمد و رفت ہوتی ہے۔  
 this cross roads through daily 20 Thousand vehicles than more  
 of arrival departure happen-pres.  
 The arrival and departure of more than 20Thousande vehicles  
 occur every day through this crossroads.

Relation-DS-r6-k2-1 :

شہریوں کو ایک بار پھر وطن لوٹ آنے کی تلقین کی۔  
 SahriyoN ko ek bAr Pir vawan lOt Ane ki talqIn kI  
 citizens Dat one time again homeland return come-Gerund -gen advice  
 do-past.  
 Citizens are once again advised to return home town.

#### 4.1.19 r6v ('kA' relation between a noun and a verb)

There are instances where a noun with 'kA' is attached to the verb but does not have any *karaka* relation. Instead, it does indicate a sense of possession. For example,

Relation-DS-r6v-1 :

شاحد کے ایک بیٹی ہے  
*SAhixa ke eka betI hE*  
 Shahid of one daughter is  
 ‘Shahid has a daughter.’

Above example is an example of a possessive sentence. Here there is a possessive relation between verb *SAhixa ke* ‘Shahid’s’ and *hE* ‘is’. The relation between this noun and the verb is marked as *r6v*.

**Syntactic cue:** In a *r6v* relation, the 'kA' vibhakti normally does not agree with the noun after it. Although, there may be cases where it does agree. In such a situation, the annotator will have to take a decision depending on the context. An example of the latter type is :

Relation-DS-r6v-2 :

نادیا 86 سال کی تھی  
*nAxiyA 86 sAla kI WI*  
 Nadiya 86 years of was  
 ‘Nadiya was 86 years old.’

In the above example, *86 varRa kI* Is predicative and is attached to the verb and not to

the noun. However, 'kI' reflects the agreement with the noun *Nadiya*.

**4.1.20 adv** (kriyAvisheSaNa 'adverbs - ONLY 'manner adverbs' have to be taken here').

Adverbs of manner are marked as 'adv'. Note that the adverbs such as place, time, etc. are not marked as 'adv' under this scheme. Place adverbs are assigned 'k7p' tag and time adverbs are marked as 'k7t'.

Relation-DS-adv-1 :

وہ جلدی جلدی لکھ رہا تھا  
vaha jalxI jalxI liKA rahA WA  
He fast fast write prog be-past  
'He was writing fast'

Relation-DS-adv-2 :

وہ بہت تیز بولتا ہے  
vaha bahuwa wejZa bolawA hE  
he very fast speak-hab be-pres  
'He speaks very fast'

**4.1.21 sent-adv** (Sentential Adverbs)

Some adverbial expressions have the entire sentence in their scope. For example,

Relation-DS-sent-adv-1 :

اس کے علاوہ بھاکپا (ماؤوا دی) کے شاہد بچن یادو کو گرفتار کر لیا گیا۔  
isake alAvA, BakaPA (mAovAxI) ke rAmabacana yAxava ko giraPZawAra kara liyA gayA  
this-of apart, BKP (maoist) of Rambacana Yadav acc arrest do take go-perf  
'Apart from this, Rambacana Yadav of BKP (Maoist) was arrested.'

In the above example, phrase 'isake alAvA' is a connective which is modifying the verb but has the entire clause in its scope. Such expressions would be attached to the verb of the sentence they are modifying and the attachment would be labeled as 'sent-adv'.

**4.1.22 rd** (relation *prati* 'direction')

The participant indicating 'direction' of the activity has to be marked as 'rd'. The label 'rd' stands for 'relation direction'.

Relation-DS-rd-1 :

نادیا گاؤں کی طرف جا رہی تھی  
nAxiyA gAzva kI waraPa jA rahI WI  
Nadiya village of direction go prog be-past  
'Nadiya was going towards her village.'

Relation-DS-rd-2 :

پیڑ کے اوپر چڑیا اڑ رہی ہے

*pedZa ke Upara ciDiyA udZa rahi hE*  
 tree of above bird fly prog be-pres  
 'The bird is flying over the tree.'

Relation-DS-rd-3:

پورب کی جانب سے میزائل داغا گیا  
*pUraba ki jAniba se missAila xAghA gayA*  
 east of direction missile launched  
 'The missile was launched from an eastern direction.'

**Syntactic cues** : An element having postpositions such as 'kI\_waraPa', ki\_jAniba etc. are to be marked as 'rd'.

#### 4.1.23 rh (hetu 'reason')

The reason or cause of an activity is to be marked as 'rh'.

Relation-DS-rh-1 :

میں نے نادیہ کی وجہ سے کتاب خریدی  
*mEne nAxiyA kI vajaha se kiwAba KArIxI*  
 I-erg Nadiya of because book bought  
 'I bought the book because of Nadiya.'

Relation-DS-rh-2 :

بارش کے سبب رکنا پڑا۔  
*bAriSa ke sababa rukana paDA*  
 Rain due to stop  
 'Had to stop due to rain.'

**Syntactic cues** : Complex postpositions such as 'ke\_sababa', 'kI\_vajaha\_se' etc are indicators of 'rh' relation. An 'rh' relation might also occur with a 'se' postposition. However, since 'se' postposition in Hindi-Urdu is highly overloaded, its presence alone can not be a deciding factor.

#### 4.1.24 rt (tadarthya 'purpose')

The purpose of an action is called as *tadarthya* which is marked as *rt*.

Relation-DS-rt-1:

میں نے نادیہ کے لیے کتاب خریدی  
*mEne nAxiyA ke liye kiwAba KArIxI*  
 I-erg Nadiya for book bought  
 'I bought the book for Nadiya.'

Relation-DS-rt-2:

میں نے جانے کے لیے ٹکٹ خریدا  
*mEne jAne ke liye tiketa KArIxI*  
 I-erg going for ticket bought  
 'I bought the ticket for going.'

Relation-DS-rt-3:

نادیا پڑھانے کے لیے اسکول جاتی ہے

*nAxiyA padZane ke liye skUla jAwI hE*

Nadiya studying for school go-hab be-Pres

‘Nadiya goes to school for studying.’

Relation-DS-rt-4:

احمد کی خاطر اکرم کو بھی دلی جانا پڑا

*ahamaxa ki khAwira akrama ko Bi xilli jAna paDA*

Ahmad for Akram Dat too Delhi go-INF

‘Akram too had to go to Delhi for Ahmad.’

Notice that in the second and third examples above, have verbs which are purpose of the action. For example in the example Relation-rt-2 *jAne ke liye* ‘for going’ is the purpose of the action *KArIdI* ‘bought’.

**Syntactic cue :** Most often 'ke\_liye' 'ke-vAswe', 'ki\_khAwira' postpositions in Urdu indicate 'rh' relation.

#### 4.1.25 ras\_k\* (upapada\_sahakArakatwa 'associative')

In sentences where two participants perform the same action but syntactically one is expressed as primary and the other as its associate, the associate participant is marked as 'ras\_k\*'. *k\** can be any karaka of which it is associative. In this tag 'r' stands for relation and 'as' stands for 'associative'.

Relation-DS-ras-1 :

شہید اپنے والد کے ساتھ بازار گیا

*SAhixa apane vAlixa ke sAWa bAjZara gayA*

Shahid own father of with market went

‘Shahid went to the market **with his father**’

Relation-DS-ras-2 :

شہید کے ساتھ نادیا نے بھی دودھ کے ساتھ کیلے کھائے

*SAhixa ke sAWa nAxiyA ne bhl xUXa ke sAWa kele KAye*

Shahid of with Nadiya erg also milk of with banana ate

‘Along with Shahid, Nadiya also ate bananas with milk’

In the first example *SAhixa* is 'k1' of the action 'gaya' (went) and since *vAlixa* ‘father’ is associative of *SAhixa* so it will be marked as 'ras\_k1'. The second example (Relation-ras-2) has two instances of *associative karakas*. 'SAhixa' is associative of 'nAxiyA', thus will be marked as 'ras\_k1' and 'k1' respectively. Also, *xUXa* ‘milk’ is associative of *kele* ‘bananas’ which is *k2* so *xUXa* will be marked as 'ras\_k2'.

**Syntactic cues :** Postposition 'ke\_sAWa' normally marks an associative relation.

#### 4.1.26 ras-NEG (Negation in Associatives)

In sentences where a karaka and its associative participate in an action but the associative does not perform the action, the associative participant is marked as 'ras-



NEG'.

Relation-DS-ras-NEG-1 :

شہاد اپنے والد کے بنا گیا  
*SAhixa apne vAlixa ke binA gayA*  
Shahid his father without went  
'Shahid went without his father'

*SAhixa* is k1 and *vAlixa ke binA* 'without his father' has an associative relationship with *SAhixa*. The relation is denoted by ras-NEG.

Relation-DS-ras-NEG-2:

احمد ساجد کے بغیر نہیں جائے گا۔  
*ahamaxa sAjixa ke baghEra nahi jAegA*  
Ahmad Sajid without not go  
'Ahmad won't go without Sajid.'

**Syntactic cues :** Postposition *ke binA* or *ke baghEra* 'without' indicates the sense of negation of associative.

#### 4.1.27 rs (relation samanadhikaran 'noun elaboration')

Elements (normally clauses) which elaborate on a noun/pronoun are annotated as 'rs'.

Relation-DS-rs-1:

بات یہ ہے کہ وہ کل نہیں آئے گا  
*bAwa yaha hE ki vo kal nahIM AyegA*  
fact this is that he tomorrow not will-come  
'The fact is that he will not come tomorrow'

*bAwa* 'fact' is 'k1' (karta) in the above example and *yaha* 'this' is its 'k1s' (k1 samanadhikaran). The relations k1 and k1s will be attached to the verb whereas the clause *ki vo kal nahIM AyegA* 'that he will not come tomorrow' will have a dependency relation with *yaha* 'this'. The relation is denoted by 'rs' (relation samanadhikaran). The main verb will take one samanadhikaran as its argument. If there are two samanadhikarans then the second samandhikaran is related with one of *karakas* with which it is associated.

Relation-DS-rs-2 :

اس نے یہ کہا کہ شہاد کل نہیں آئے گا  
*usane yaha kahA ki SAhixa kala nahIM AyegA*  
he-erg this told that Shahid tomorrow not will-come  
'He told that Shahid will not come tomorrow.'

In Relation-DS-rs-2 above the complement clause is the complement of the karma pronoun *yaha* 'this'. Therefore, it will be attached to the pronoun 'yaha' and would also be labeled as 'rs'. While annotating the sentence, the conjunct 'ki' will be annotated as 'rs' will be attached to the 'yaha' which is the k2 of the verb of the main clause ('kahA' in this case). The finite verb of the complement clause ('nahIM AyegA' in the above example) will be attached to the conjunct 'ki' (that) and would be labeled

as 'ccof'.

#### 4.1.28 rsp (relation for duratives)

The durative expressions have two points – a point of starting and an end point. The expression as a whole may express time, place or manner etc. The tag 'rsp' shows the relation between the starting point and the end point of a durative expression. For example,

Relation-DS-rsp-1 :

۱۹۹۰ سے لے کر ۲۰۰۰ تک ہندوستان کی ترقی تیز رہی

*1990 se lekara 2000 waka hinxuswAna kl waraqqi  
wejZa rahl*

1990 from taking 2000 till India of development fast was  
'India was fast developing from 1990 till 2000'

The entire expression *1990 se lekara 2000 waka* 'from 1990 till 2000 a time expression. There are two parts in this time expression, one is starting point(1990) the other is the ending point(2000). The vibhaktis *se* 'from' and *waka* 'till' give us the information of starting point and ending point in time. As the entire expression *1990 se lekara 2000 waka* is a time expression it will have a k7t (time relation) relation with the verb. Now internally the two parts of the time expressions are related to each other. So the relation of *1990 se 2000* 'from 1990' with *2000 waka* 'till today' will be rsp (relation source of a durative).

**Syntactic cues :** Duratives will have 'se lekara - - - waka' construction.

#### 4.1.29 rad (address terms)

Terms such as *janAba*, *mohwarama* etc. are the address terms. Such terms are annotated as 'rad'.

Relation-DS-rad-1 :

جناب، مجھے کل دلی جانا ہے

*janAba, muJe kala xillI jAnA hE'*

Sir, I-Dat tomorrow Delhi to go be-pres

'Sir, I have to go to Delhi tomorrow'

Relation-DS-rad-2 :

ماسٹر صاحب، کیا کل اسکول کھلا ہے

*mAstara sAhaba, kyA kala skUla KulA hE*

master hon what tomorrow school open be-Pres

'Teacher, is the school open tomorrow?'

#### 4.1.30 nmod\_\_relc, jjmod\_\_relc, rbmod\_\_relc (relative clauses, jo-vo constructions)

A relative clause construction in Hindi-Urdu has a 'jo' pronoun. Typically, the modified element has a pronoun 'vaha' in it. Such relative clauses where there is a corresponding 'vaha' pronoun in the main clause are called relative-correlative (*jo-vo*)

constructions. The *jo-vo* constructions in Hindi-Urdu are highly productive. These occur not only as noun modifiers but also as modifiers of adjectives and manner adverbs.

Relative\_clause-DS-1 :

میری بہن جو دلی میں رہتی ہے کل آ رہی ہے  
*merI bahana [jo xillI meM rahawI hE] kala A rahI hE*  
 my sister who Delhi in live-hab pres tomorrow come prog pres  
 'My sister who lives in Delhi is coming tomorrow'

The above example does not have a 'vaha' pronoun in the modified NP. Relative clauses without a 'vaha' pronoun in the modified NP normally are elaborative in nature. These are also not so frequent.

A relative clause can be either prenominal or postnominal.

**(a) Prenominal:** The relative clause occurs to the left of the head noun and it carries a relative pronoun 'jo' as a demonstrative along with the noun. For example,

Relative\_clause-DS-2:

جو لڑکا وہاں کھڑا ہے وہ میرا بھائی ہے  
*[jo ladZakA vahAz KadZA hE] [vaha merA BAI hE]*  
 who boy there standing pres he my brother is  
 'The boy who is standing there is my brother'

Relative clause in the above example is modifying 'vaha' of the main clause. However, 'vaha' itself refers to 'ladZakA' which occurs in the subordinate relative clause along with the relative pronoun 'jo'. Thus, the relative clause has '*jo ladZakA*' as the relativizing element. The pronoun *vaha* 'he' in the main clause has '*jo ladZakA*' as its referent. The prenominal relative clauses in Hindi-Urdu mostly have this structure.

**(b) Postnominal:** The relative clause occurs to the right of the head noun and the relative pronoun in such cases behaves like a full-fledged pronoun and is not a demonstrative any more.

Relative\_clause-DS-3 :

وہ لڑکا جو وہاں کھڑا ہے میرا بھائی ہے  
*vaha ladZaka [jo vahAz KadZA hE] merA BAI hE*  
 that boy who there standing pres my brother is  
 'The boy who is standing there is my brother'

A relative clause can also occur to the right of the main verb as in the following example:

Relative\_clause-DS-4 :

وہ لڑکا میرا بھائی ہے جو وہاں کھڑا ہے  
*vaha ladZakA merA BAI hE [jo vahAz KadZA hE]*  
 that boy my brother is who there standing pres  
 'That boy is my brother who is standing there.'

A relative clause can modify any element in the main clause whatever participatory role it might have. Thus a relative clause can modify a *karta* (subject/agent), *karma* (direct object), *samradana* (indirect object), *karana*, *adhikarana* (oblique object) etc.

**(i) *karta* (subject) modification :**

Relative\_clause-DS-5 :

جو لڑکا وہاں بیٹھا ہے وہ میرے اسکول میں پڑھتا ہے  
*jo ladZakA vahAz bETA hE vaha mere skUla meM paDZawA hE*  
 who boy there sit\_perf is he my school in study\_hab pres  
 'The boy who is sitting there studies in my school.'

**(ii) *karma* (object) modification) :**

Relative\_clause-DS-6 :

شاحد نے وہ سیب کھایا جو خراب ہو گیا تھا  
*SAhixa ne vaha seba KAyA jo KArAba ho gayA WA*  
 Shahid erg that apple ate which rotten happen go-perf be-past  
 'Shahid ate an apple which was rotten.'

**(iii) *sampradana* (Indirect object) modification :**

Relative\_clause-DS-7 :

شاحد نے اس لڑکے کو کتاب دی جو وہاں کھڑا تھا  
*SAhixa ne usa ladZake ko kiwAba xI jo vahAz KadZA WA*  
 Shahid erg that boy acc book gave who there standing be-past  
 'Shahid gave the book to that boy who was standing there.'

**(iv) *karana* (Oblique object) modification :-**

Relative\_clause-DS-8 :

شاحد نے اس چاقو سے سیب کاٹا جو تیز تھا  
*SAhixa ne usa cAkU se seba kAta jo wejza WA*  
 Shahid erg that knife by apple cut which sharp was  
 'Shahid cut an apple with the knife which was very sharp.'

Given below are the examples and corresponding tags for the 'jo-vo' constructions of Hindi-Urdu :

**a. nmod\_\_relc (relative clause constructions modifying a noun)**

Relation-DS-nmod\_\_relc-1 :

جو لڑکا وہاں بیٹھا ہے، وہ میرا بھائی ہے  
*jo ladZakA vahAz bETA hE, vaha merA BAI hE*  
 who boy there sat is he my brother be-pres  
 'The boy who is sitting there is my brother.'

Since it is an entire clause which modifies a element in the main clause, the

convention which is followed in the current annotation scheme is to attach the verb of the subordinate clause to the element it modifies. The relation between 'jo' and 'vo' is marked by showing a co-referential tag (coref). Therefore, a tree representation for the above example would be as follows:

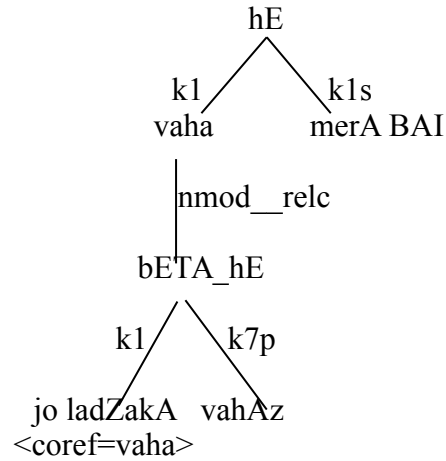


Figure 3

#### b. **rbmod\_\_relc** ('jo' construction modifying an adverb)

A relative-corerelative construction can occur for an adverbial expression as well. Such 'jo' clauses would be attached under the adverb they modify with a tag 'rbmod\_\_relc'.

Relation-DS-rbmod\_\_relc-1:

شاحد نے جیسا کیا، میں نے بھی ویسا ہی کیا  
*SAhixa ne jEsA kiyA, mEMne BI vaisA hI kiyA*  
 Shahid erg like-what did, I-erg also like-that emph did  
 'I did exactly what Shahid did.'

#### c. **jjmod\_\_relc** ('jo' construction modifying an adjective)

A 'jo' clause can also modify an adjective. It will be annotated as **jjmod\_\_relc**

Relation-DS-jjmod\_\_relc-1 :

مکان ویسا ہی خوبصورت بناؤ جیسا کہا گیا ہے  
*makAna vEsA hI khUbsUrawa banAo, jEsA kahA gayA hE*  
 house like-that part.beautiful build like-what told go-perf pres  
 'Build a house as beautiful as has been told'

(Here the clause containing **jEsA** is modifying the adjective **vEsA sunxara** )

#### 4.1.31 **nmod** (participles etc modifying nouns)

**nmod** is an underspecified relation employed to show general noun modification without going into a finer type. Since the dependency relations are being marked at the chunk level, simple adjective modifiers do not normally occur at this level. An adjective noun sequence is already chunked and their dependency relations are

marked only when the chunks are expanded into dependency sub-trees. A tag 'adj' is used for marking simple adjective – noun modification. This tag is not discussed in this document. The nominal modification by adjectival participles falls within the purview of this document. However, an underspecified tag 'nmod' is used to show these dependencies.

Relation-DS-nmod-1 :

پیڑ پر بیٹھی چڑیا گانا گارہی تھی  
*pedZa para bETI cidZiyA gAnA gA rahI WI*  
 tree on sitting bird song sing prog be-past  
 'The bird sitting on the tree was singing a song.'

In the above example, the participle clause 'pedZa para bETI' is modifying the noun 'cidZiyA'. Following a tree representation of the above sentence:

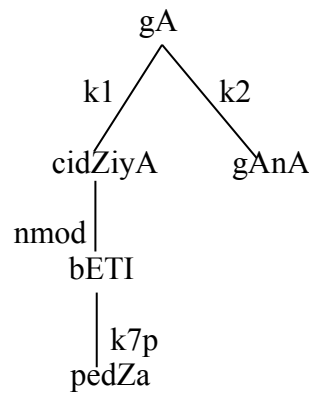


Figure 4

**Syntactic cues :** The non-finite verb form of such participial modifiers agree in gender and number with the noun it modifies. The gender and number of the verb 'bETI' in the above example agrees with the gender and number of the noun 'cidZiyA'.

#### 4.1.32 vmod (verb modifier)

'vmod' is another underspecified tag. For some relations getting into finer subtypes is not yet possible. Such relations are annotated with slightly underspecified tag, a tag high on the dependency tag type tree given in Figure 2 under section 3.2.3. 'vmod' is one such tag. A verb (especially non-finite) that modifies another verb is thus marked as 'vmod'. There can be two types of verb modifiers:

**(a) Simultaneous :** where the actions denoted by the two verbs modifier and modified happen simultaneously.

Relation-DS-vmod-1 :

وہ کھاتے ہوئے گیا  
*vaha KAwe hue gayA*  
 he eat-Impf-prtpl went  
 'He left while eating'

**(b) Sequential :** where one action happens after the completion of the another action.

Relation-DS-vmod-2 :

وہ کھانا کھا کر گیا

vaha *KAnA KAkara* gayA  
he food having-eaten went  
'He left after eating the meal'

Relation-DS-vmod-3 :

اس کو وہاں گئے ہوئے کئی دن بیت گئے ہیں

usako vahAM gaye hue kaI xina bIwa gaye hEM  
he-Dat there go-perf prtpl several days pass go-perf be  
'A number of days have passed since he went there.'

**(c) '-kara' participles in Hindi-Urdu:** Most Indian languages have a high frequency of participial usages. So does Urdu. Of various participles in Hindi-Urdu, 'kara' is one of the most frequent one. It also serves several semantic functions. One of them is showing sequentiality of events (example Relation-vmod-2 above). Other than *sequential*, *kara* participle has other senses also. They are:

**(i) Consequential :** In case of a 'kara' participle modifying another verb, the 'kara' participle expresses the causality of the other action.

Consequential\_kara-DS-1 :

شہید سانپ کو دیکھ کر ڈر گیا

SAhixa sAzpa ko *xeKakara* dara gayA.  
Shahid snake acc having seen fear go-past  
'Having seen the snake Shahid got frightened.'

**(ii) Manner :** The 'kara' participle in certain cases expresses the manner of the verb it modifies.

Manner\_kara-DS-1:

شہید بھاگ کر آیا

SAhixa *BAgakara* AyA.  
Shahid running came  
'Shahid came running.'

**(iii) Instrument :** 'kara' participle also acts as an instrument of the verb it modifies.

Instrument\_kara-DS-1:

شہید محنت کر کے پیسے کماتا ہے

SAhixa mehanawa *karake* pEse kamAwA hE.  
Shahid hard-work having done money earn be-res  
'Shahid earns money by working hard.'

All the above constructions with *kara* and *wA huA* are vmods. Finer analysis for the above is done. However, it has been decided to mark all of the above as 'vmod' only.

#### 4.1.33 jjmod (modifiers of the adjectives)

The tag for modifiers of the adjective is also an underspecified tag. In this case finer relations have not been worked out as yet since the need for finer relation tag for adjective modifiers is not felt for syntactic annotation. Therefore, the tag for marking adjective modifiers is 'jjmod'.

Relation-DS-jjmod-1:

ہلکی نیلی کتاب  
*halkI nIII kiwAba*  
*light blue book*  
'Light blue book'

( The word *halkI* 'light' in the above example is modifying the adjective *nIII* 'blue' and not the noun *kiwAba* 'book')

#### 4.1.34 pof (part of units such as conjunct verbs)

A conjunct verb is a verb that is formed by combining a noun or an adjective with a verb. Therefore, the internal structure of a conjunct verb would be [noun/adj + verbalizer]. Conjunct verbs are highly productive in Hindi-Urdu. 'karana, honA' are the most commonly occurring verbalizers in Hindi-Urdu. Some of the other verbalizers are 'lenA, denA'. Identifying a conjunct verb is a difficult process in Hindi-Urdu as the syntactics diagnostic tests work only upto a point and not beyond. Literature on the definite syntactic behaviour of conjunct verbs does suggest a number of diagnostics though (Mohanani 1994; Butt, 2004; Chakrabarty et. al, 2007; Bhatt, 2008).

In the current scheme a special tag 'pof' has been introduced to mark the conjunct verbs. 'pof' does not exactly denote a dependency. It rather represents that the two elements related by this tag are part of a multi word expression (MWE). Therefore, the relation between the two elements of the conjunct verb **ghusla + karana** 'bath + do' would be shown as follows :

kara  
|  
pof  
|  
ghusla

Figure 5

Some examples of conjunct verb constructions are given below :

Relation-DS-pof-1:

شہید سیف کا انتظار کر رہا تھا  
*SAhixa sAifa kA inwezAra kara rahA WA.*  
Shahid Saif of wait do prog be-past  
'Shahid was waiting for Saif'

Relation-DS-pof-2:



شاحد نے ایک سوال کیا

*SAhixa ne eka savAla kiyA*

Shahid erg one question did

'Shahid asked a question'

In Relation-DS-pof-1, *inwezAra kara* 'to wait' is a conjunct verb. The relationship between *inwezAra* and the verb *kara* 'do' will be marked as *pof*. In the second example above *savAla kiyA* 'questioned' is a conjunct verb. But *savaal* 'question' has a modifier *eka* 'one'. The issue here is – semantically 'savAla karana' is one unit. Therefore, it is logical to group them together within a verb chunk. However, since the noun of a conjunct verb retains its nominal property and can be modified by an adjective (example Relation-DS-pof-2 above), we should be able to represent it in the dependency tree. Grouping them together within a verb chunk would fail to address the problem of an element modifying the noun element of a conjunct verb. *eka* 'one' in the above example is a modifier of *savAla*. *savAla* itself is a part of the conjunct verb *savAla kiyA*. Since *savAla kiyA* is already grouped as one chunk, it is not possible to establish relation between *eka* and *savAla*. Therefore, the noun 'savAla' would be chunked separately from the verb 'kiyA' (Bharati et al., 2006). However, the fact of 'savAla' and 'kiyA' being parts of a single unit, a conjunct verb, needs to be captured.

To overcome this problem it was decided that we tag the noun of the conjunct verb as NN at the POS level. Thereafter, the noun is grouped with its preceding adjectival modifiers (if any) 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 at the chunk level and the noun of the 'conjunct verb' is separated from its verbalizer. Thus, we show the 'parts-of' relation between the noun and the verbalizer of a conjunct verb, using *pof* tag.

The advantage of this solution is that:

- 1) It allows us to show the modifier-modified relation between an adjective such as *eka* 'one' in the above example with its modified noun *savAla* 'question'.
- 2) Since the information of a noun verb sequence being a 'conjunct verb' is crucial at the syntactic level, it is captured at this level by marking the relation between the 'noun' and its verbalizer by an appropriate tag.

As mentioned above there are problems in identifying conjunct verbs in a sentence in Hindi-Urdu. The available syntactic tests (Mohan 1994; Chakrabarty et. al, 2007; Bhatt, 2008) are not very satisfactory. This appears to be an issue for syntax – semantic interface. There are several cases where a native speaker is quite convinced that a noun verb sequence is a case of conjunct verbs. However, syntactically the noun behaves more like an argument of the verb. In the absence of satisfactory tests for identifying a conjunct verb, several noun verb sequences pose a major problem for the annotators on whether to treat them as conjunct verbs or otherwise.

Therefore, as of now, the decision has been left to the annotators with a full understanding that this may lead to some inconsistency in the data. The final decision of when a noun verb sequence is a conjunct verb and when not has been left to the senior linguists who would do some checks on the annotated data. Given below are a

number of examples of Hindi-Urdu conjunct verbs :

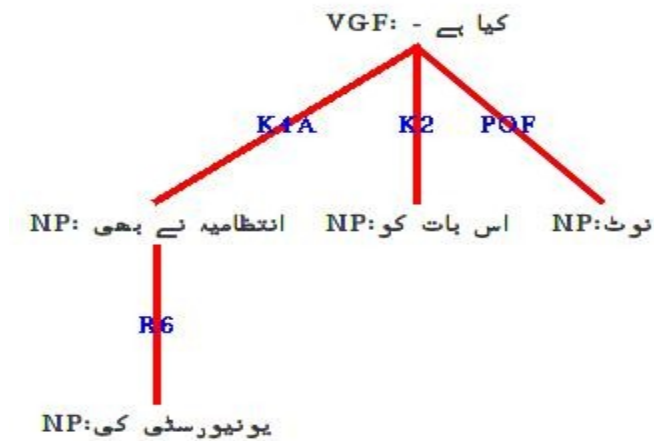
(.jpgk4a) یونیورسٹی کی انتظامیہ نے بھی اس بات کو نوٹ کیا ہے۔ 1

yunivrsty ky intzamyh bhi is bat ko not kya he.

university ki intezamiya ne bhi is baat ko note kiya hae.

university of management [ne(erg.)]too this point [ko(Dat.)]note do-past be-pres.

the management of the University too has noted this point.



(.jpgpof) کئی کنٹل چاول کو بھی ضبط کر لیا گیا ہے۔

kyi qntl chavl ko bhy zbt kr lya gya he.

kayi qintal chaval ko bhi zabt kar liya gaya hae.

several quintal rice to[ko-Dat.]too seize do take passive be-pres.

several quintals of rice has been seized.



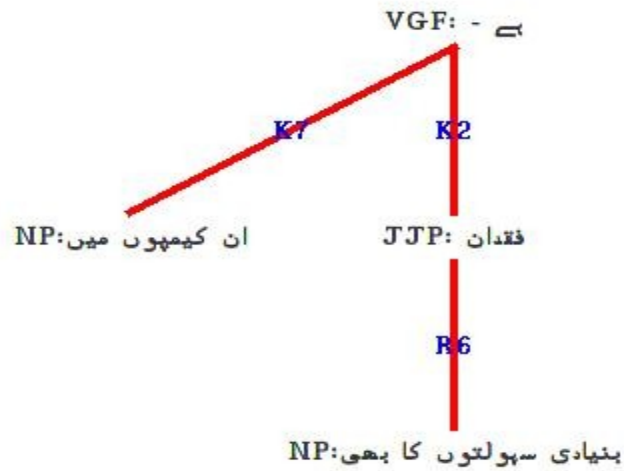
2 (r6.jpg) ان کیمنوں میں بنیادی سہولتوں کا بھی فقدان ہے۔

in kympon myn bnyady sHulton ka bhy fqdan he.

in kamiyon me bunyadi sahulaton ka bhi faqdan hae.

these camps in basic facilities of too absence is(be-pres).

these camps are even short of basic aministrations.



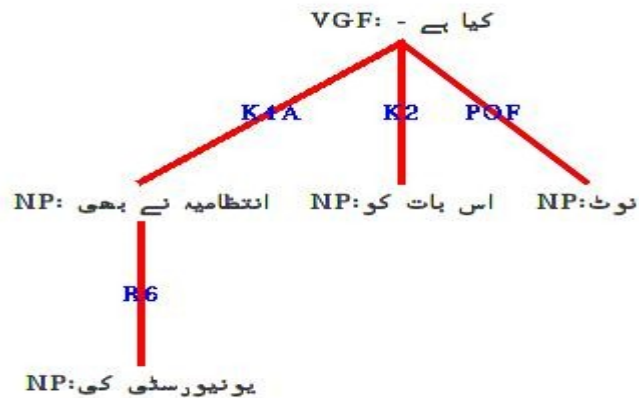
1 (jpgk4a) یونیورسٹی کی انتظامیہ نے بھی اس بات کو نوٹ کیا ہے۔

yunivrsty ky intzamyh bhi is bat ko not kya he.

university ki inteزامiya ne bhi is baat ko note kiya hae.

university of management [ne(erg.)]too this point [ko(Dat.)]note do-past be-pres.

the management of the University too has noted this point.



**B.** Since 'pof' indicates a 'part of' relation between two words of a single lexeme, it is generalized to indicate relation between different elements of other MWEs as well. Hence in the following example, 'PulA nahIM samAyA' is an idiom and 'pof' will be used to mark the relation between 'PulA' and 'nahIM samAyA'.

Relation-DS-pof-7 :

شہاد خوشی سے پھولا نہیں سمایا  
*SAhixa KuSI se PULa nahIM samAyA*  
 Shahid happiness because of bloated not contained  
 'Shahid was bursting with happiness.'

Label 'pof' has three subtypes :

- (1) pof (conjunct verb)
- (2) pof-idiom (idiom)
- (3) pof-compound (compound noun)

Example (Relation-DS-pof-7) has an idiom *PulA nahIM samAyA* 'was bursting with happiness', the parts of this idiom would be connected by the label 'pof'.

#### 4.1.35 ccof (co-ordination and sub-ordination)

Another special tag which does not exactly reflects a dependency relation is 'ccof'. This is used for coordinating as well as subordinating conjunctions. The dependency trees will show the conjuncts as heads. In case of coordinating conjuncts, the conjunct is the head and takes the coordinating elements as its children. Likewise, a subordinating conjunct would take the clause to which it is syntactically attached (the subordinate clause) as its child.

#### (a) co-ordinating conjunct :

Relation-DS-ccof-1:

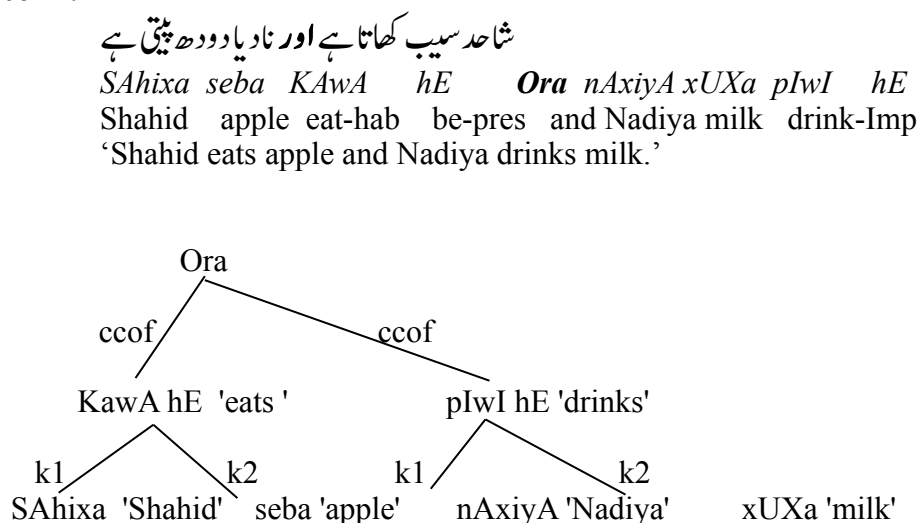


Figure 6

The above example is an example of co-ordination of two clauses. However, the tag

'ccof' would be used for any co-ordination. Therefore, co-ordination of nouns, adjectives or adverbs will all be tagged with a 'ccof' tag. Following is an example of noun co-ordination :

Relation-DS-ccof-2:

شاحد اور حامد اسکول جاتے ہیں

*SAhixa Ora hAmixa skUla jAwe hEM*

Shahid and Hamid school go-hab be-pres

'Shahid and Hamid go to school.'

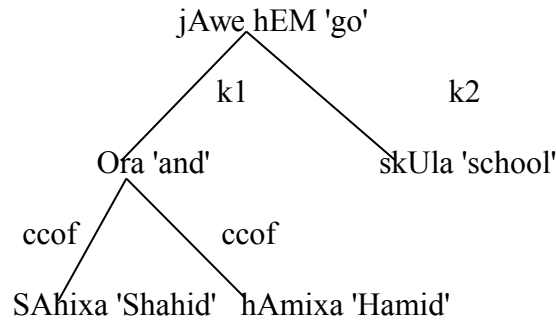


Figure 7

Relation-DS-ccof-3 :

شہروں کی سڑکوں پر شام و سحر بھیڑ لگی رہتی ہے۔

*Sahara ki saRakoN para Sama o sahara BID lagI rahawI hai*

city of roads on evening and morning crowd is

'The city's roads are crowded all morning and evening.'

(b) sub-ordinating conjunct :

Relation-DS-ccof-2 :

شاحد نے حامد سے کہا کہ وہ کل نہیں آئے گا

*SAhixa ne hAmixa se kahA ki vaha kala nahIM AyegA*

Shahid erg Hamid to told that he tomorrow not will-come

'Shahid told Hamid that he will not come tomorrow.'

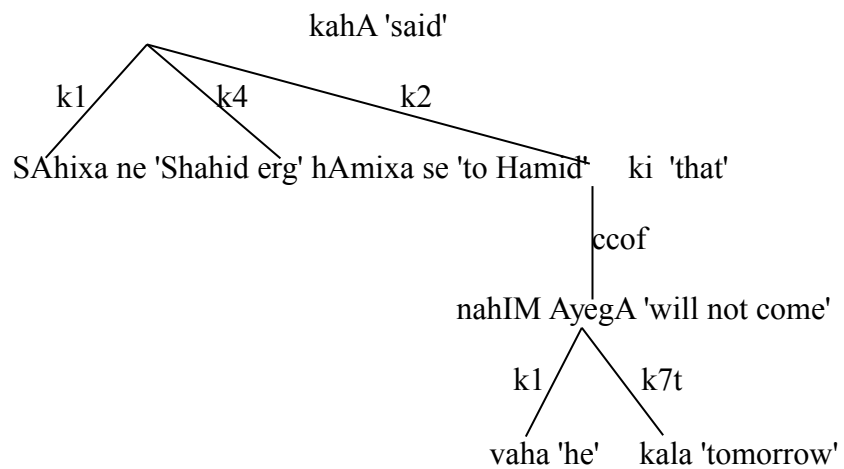


Figure 8

A coordinating conjunct would have two or more branches which would be labeled as 'ccof' and a subordinating conjunct would have only one branch.

#### 4.1.36 fragof (Fragment of)

'fragof' is a tag which has been included to handle some very special cases.

A. There are examples in the corpus where a postposition, a negative particle or an auxiliary are separated from the NP or VP of which normally they are a part of. Thus, they do not occur as part of the chunk where they belong. For example,

Relation-DS-fragof-1 :

بھاکپا (ماؤوا دی) کے را مہجین یادو کو گرفتار کر لیا گیا

**BakaPA** (*mAovAxI*) **ke** *rAmabacana yAxava ko giraPZawAra kara liyA gayA*  
 BKP (maoist) of Rambacana Yadav acc arrest do reflx-perf go-perf  
 'Rambacana Yadav of BKP (Maoist) has been arrested.'

In the above example, the NP chunk '*BakapA ke*' has been broken through the insertion of additional information (*mAovAxI*) about '*BakapA*'. The noun '*mAovAxI*' itself forms a separate NP chunk. Therefore, the expression **BakaPA** (*mAovAxI*) **ke** would appear as follows in chunks :

```
((      NP
  BApA      NNP
))
((      NP
  (        SYM
  mAovAxI   NN
  )        SYM
  ))
((      FRAGP
  ke      PSP
  ))
```

SSF-2

The expression '*BakapA ke*' is broken into two chunks. The postposition '*ke*' which is separated from its noun '*BakapA*' is chunked as 'FRAGP'. To represent that the post position '*ke*' is part of the noun chunk '*BakapA*', the postposition chunk would be annotated with the value 'fragof' for the attribute 'drel'.

This is a tag which is an exception in the normal scheme as it marks the relation of two members of the same chunk. Also, this chunk would normally contain a function word which is a part of some other chunk. After annotating the value 'fragof' for the attribute 'drel', the FRAGP chunk would appear as follows :

```
((      FRAGP      <drel=fragof:NP>
  ke      PSP
  ))
```

SSF-3

The occurrence of such cases could be due to some intervening material or some time

the main part of the chunk is dropped.

**B.** There are also instances where the main part of the chunk is missing. It normally happens in cases of gapping particularly with negative particles.

Relation-DS-fragof-2 :

بہار کے گورنر کو نوٹس بھیجا جاسکتا ہے کہ نہیں  
*bihAra ke governor ko notisa BejA jA sakawA hE ki nahIM*  
 Bihar of governor acc notice send go can is or not  
 'Can the notice be sent to the Bihar Governor or not ?'

In the above example, the second occurrence of the verb 'BejA' has been omitted. Consequently, only the negative particle 'nahIM' is left. To represent the dependencies of the second clause, it is important to insert a verb node. Since, in the current scheme, the negative particles are chunked with the verb, this intra-chunk relation would then be represented by marking the negative particle with 'fragof'. Therefore, the verb chunk and the negation chunk would appear as follows after annotation :

```
((      NEGP  <drel=fragof:NULL__VGF>
nahIM
))
((      NULL__VGF  <name=NULL__VGF>
NULL VG
))
```

SSF-4

#### 4.1.37 enm (enumerator)

The tag 'enm' is another special tag. This tag also does not represent a dependency in the strict sense. Although, this again is a value for the attribute 'drel' of the word. This tag is used to mark the enumerators such as 1, 2, 3 or a, b, c, etc in a text. These enumerators occur in the beginning of a sentence and they need to be attached to the root node. In the treebank, the root node normally, is either a verb or a conjunct. Therefore, it has been decided to attach the enumerators to the verb with a label 'enm'. For example,

Relation-DS-enm-1

*1. Apa apanA tax samaya se xe sakawe hEM*  
 1. you your tax time on give can be-pres  
 '1. You can pay your taxes on time.'

In the above example, numeral '1.' has occurred as an enumerator. This will be chunked separately with a chunkd label 'BLK'. At the dependency level, this chunk will be attached to the verb 'xe sakawe hEM'. Therefore, the annotated example would be :

```
((      BLK  <drel=enm:VGF>
1      QC
.      SYM
))
SSF-5
```

## 4.2. How to Mark Elided Elements ?

An issue that came up before us while working on the scheme was whether to mark elided elements in a sentence or not. After due deliberations, it was decided to mark a missing element in the tree for the following cases :

- (a) In case of a missing verb since a verb forms the root node of a tree/subtree (see section on Gapping (4.2.1) for more details)
- (b) In case of a missing co-ordinating conjunct since it also forms the root of a co-ordinating tree under the current scheme.
- (c) In case of any other node which may be a root node for a tree or a sub-tree. For example, 'ulleKanIya hE ki .....,
- (d) In case of missing arguments of a verb. Amongst the missing arguments, it was decided to mark only k1 and k2. However, The missing arguments will be inserted only in the following cases:
  - (i) Shared arguments
  - (ii) Gapping
  - (iii) Also in finite subordinate clauses

For making the above missing elements explicit it was decided to introduce a NULL node in the tree. The node would be chunked and the relevant features would be annotated at the chunk level depending on the type of the node inserted. The details of the features to be annotated for various types have been provided under the cases discussed below.

In the following sub-sections each of the above, except 'shared arguments', is discussed in more details. The shared arguments have been discussed in more details under Section 4.3. below.

### 4.2.1 Gapping

Gapping is a type of ellipses where a verb is omitted in its repeat occurrences. Some times the arguments of the verb may also be omitted along with the verb. Ross (1967) introduced the term. An example of gapping in Hindi-Urdu is given below :

Gapping-DS-1:

شاحد دلی گیا اور سیف آگرا

*SAhixa xillI gayA Ora sAifa AgarA*

Shahid Delhi went and Saif Agra

'Shahid went to Delhi and Saif to Agra.'

In the above example the occurrence of the verb 'gayA' (went) in the second clause of the co-ordinating construction has been elided. To complete the dependencies of the second clause, it is essential to explicitly show the verb which would be the root node of the tree. The missing verb can be retrieved from the previous clause. Thus, the gapped element would be marked as follows :

- (i) First a new node would be created :  
NULL                  VM



No other information about this node would be provided.

(ii) Next, the above node would be chunked. The chunk would be annotated for the following features :

<name=" troot=" mtype=">

Of the three attributes given above, 'name' is an attribute which is annotated on all chunk nodes. The attribute 'troot' is to be added for a gapped verb as it is retrievable from the context. The attribute 'mtype' is to mark every missing element for whether it is a case of 'gap' or 'not'. Therefore, this attribute would have only two values (1) gap and (2) non-gap.

In case the gapped verb is also a dependent of a higher node, an additional attribute of 'dmrel' would be annotated as well. The attribute 'dmrel' is same as 'drel'. The attribute 'drel' is for the words in a sentence and the attribute 'dmrel' would be on elements which are not present in the sentence explicitly. Thus, the chunk annotated for the gapped element in the above example would look as follows:

```
((      NULL__VGF      <name='NULL__CCP' troot='jA' mtype='gap'>
NULL VM
))
```

SSF-6

The example below is another case of gapping.

Gapping-DS-2 :

شاحد نے نادیا کو کتاب دی اور عاطف نے مریم کو  
*SAhixa ne nAxiyA ko kiwAba xI Ora AwIPZa ne mariyama ko*  
Shahid Erg Nadiya acc. book gave and Atifa Erg Maryam acc.  
‘Shahid gave a book to Nadiya and Atif to Tina.

However, in the above example, an argument is also dropped in the second clause. This argument and the verb can be retrieved from the previous clause. To build a complete dependency tree for the above example, the following items will be inserted in the tree, (a) the missing verb and (b) the missing argument.

The following chunks for (a) and (b) will be created respectively :

```
((      NULL__VGF      <troot='xe' name='NULL__VGF' mtype='gap'>
NULL VM
))
((      NULL__NP      <dmrel=k2:xe reftype=cotype:kwAba name='NULL__CCP
mtype='gap'>
NULL NN
))
```

SSF-7

#### 4.2.2 Missing co-ordinating conjunct

Some times the co-ordinating conjunct is implicit and does not occur in the sentence explicitly. For example,

Elided-conjunct-DS-1:

بچے بڑے ہو گئے ہیں کسی کی بات نہیں مانتے  
*bacce badZe Ho gaye hEM kisI kI bAwa nahIM mAnawe*  
children big happen go-perf be-pres no-one's of talk not listen to  
'The children have grown big and do not listen to anyone.'

In the above example, the co-ordinator 'Ora' is missing. Since co-ordinating conjunct forms the root node, a NULL node will be inserted to represent it. Thus, the example after the insertion of NULL would appear as:

Elided-conjunct-DS-1:

*bacce badZe Ho gaye hEM NULL kisI kI bAwa nahIM mAnawe*

The feature structure for the NULL node would be :

```
((      NULL__CCP  <name=NULL__CCP>
  NULL  CC
))
SSF-8
```

#### 4.2.3 Missing root node

A commonly occurring construction in Urdu is :

Missing-yaha-DS-1:

*vAze ho ki unhoMne yaha bAwa mAna II*  
be-it-known that they this suggestion accept reflex-past  
'It should be known that they accepted this proposal.'

In the above example, the sentence begins with an adjective and has a complement clause in the predicative position. The highlighted words show the adjective, verb be and the complement 'ki'. The complement clause in such sentences is actually an NP complement of the subject, which is missing. To represent this a NULL node is to be inserted and the clause is can then be attached to it as its modifier. The inserted NULL node in this case would look like :

```
((      NULL__NP  <name=NULL__NP troot=yaha mtype=non-gap>
  NULL  NN
))
SSF-9
```

#### 4.2.4 Missing arguments in a co-ordinating construction :

The example Gapping-DS-2 above shows a case of an elided argument along with the gapped verb. In case of gapping, the verb is same in both the clauses and consequently its repeat occurrence is omitted. It is also possible that the two clauses in a co-ordinate structure may have two different verbs. In such a situation both the

verbs are realized explicitly. However, the repeated arguments in a co-ordinated construction are dropped even if the verb is different and is realized on surface. For example,

Elided-arg-DS-1 :

نادیا نے کتاب پڑھی اور سو گئی  
*Nadiya ne kiwAba padZi Ora so gayI*  
 Nadiya Erg book read and sleep go-Past  
 'Nadiya read the book and slept.'

In the above case both the verbs '*padZI*' (read) and '*so gayI*' (slept) have Nadiya as their *karta* (k1). However, the second occurrence of Nadiya is omitted. In such cases also, the missing argument would be inserted and would be represented as follows:

```
((      NULL__NP <name=NULL__NP mtype='gap' dmrel='k1:VGF2'
reftype=corefn:nadiya>
  NULL NN
))
SSF-10
```

#### 4.2.5 How to represent missing co-ordinated constructions

There are cases where a verb has a missing argument which from the context appears to be a co-ordinated construction. The issue then is how to represent it in our tree and what all to insert? For example,

Elided-arg-DS-2:

میرے بھائی شاحد نے اور اس کے دوست سیف نے باہر کھانا بھی کھایا اور پکچر بھی دیکھی  
*mere BAI SAhixa ne Ora usake xoswa SAifa ne bAhara KAnA BI KayA Ora pikcara BI xeKI*  
 my brother Shahid erg and his friend Saif erg outside food also ate and movie also saw  
 'My brother Shahid and his friend Saif had a meal outside and also watched a movie.'

The example above has two verbs *KayA* 'ate' and *xeKI* 'saw'. The *karta* of *xeKI* 'saw' is same as that of *KayA* 'ate'. This *karta* is a co-ordinated NP (*mere BAI SAhixa ne Ora usake xoswa SAifa ne*) and is elided in its second occurrence. Thus *xeKI* 'saw' does not have an overt *karta*. Since we have decided to mark all missing k1, we have to then insert a NULL element and show its co-reference to the explicitly present argument of the previous verb *KayA* 'ate'.

In such cases of co-ordinated missing arguments a NULL CCP would be inserted and it would be co-referred to the explicitly present CCP *karta*. Therefore the SSF representation of the inserted NULL\_\_CCP would be as follows:

```
(( NULL__CCP <name=NULL__CCP dmrel=k1:VGF2 mtype=gap
reftype=corefn:CCP>
  NULL CC
))
(( VGF <name=VGF2>
  xeKI VM
))
SSF-11
```

### 4.3 How to mark shared arguments ?

Since Hindi-Urdu allow omitting of mandatory arguments, there are a number of sentences with missing arguments. Missing arguments in a sentences could be due to being shared between two or more verbs or due to ellipsis. The difference between sharing and omitting is that in sharing the argument occurs once which is shared by two verbs ie. main verb which would be finite and the participle clause which would have a non-finite verb. In sharing the second argument can not be realized syntactically. The other case of missing argument is when the argument can (in principle) occur twice but it has been dropped in the second clause (as in case of gapping).

Since k1 and k2 are otherwise mandatory arguments for several verbs and these two arguments also play a crucial role in several linguistic decisions, it was decided to make them explicit in case they were missing in a sentence. For making the missing k1 and k2 explicit the following procedure has to be followed.

- a) Insert a NULL node in the tree for a missing argument.
- b) Assign it appropriate POS tag, normally a NN.
- c) Chunk the NULL node and assign it appropriate chunk label. However, it has to be prefixed with NULL\_\_ . As shown above (in 4.1), the label for missing verb chunk would be 'NULL\_\_VGF'. For a missing nominal argument, it would be 'NULL\_\_NP'.
- d) As mentioned earlier, a new dependency attribute is introduced in the scheme to mark the dependency relations of the inserted nodes. The attribute is 'dmrel'. 'dmrel' stands for 'dependency relation for a missing element'.
- e) Missing argument could either be co-referential with another element in the tree or could be of the same type but not exactly co-referential. Thus, to mark this distinction an attribute 'reftype' has been introduced. The values for the 'reftype' would be 'corefn:X' or 'cotype:X'. The value has three parts to it. The first part (corefn, cotype) indicates the 'type' of reference, the second part (:) indicates 'of' and the third part 'X' stands for 'what'. Please see example under section on shared argument for more clarity.

Therefore, the following information is annotated in an inserted node for a missing argument :

```
((      NULL__NP    <name='NULL__NP' dmrel=" reftype=" mtype=">
  NULL NN
))
```

*SSF-12*

NOTE : The attribute 'troot' is not annotated for a missing argument as it is captured by the 'reftype'. In principle, the morph features (root, number, gender, person) of the corresponding element in the sentence can be copied to the inserted node and need not be manually annotated.

Coming back to the sharing of arguments, the sharing of arguments can be of two types :

#### 4.3.1 Sharing in non-adjectival participles:

In non-adjectival participles, an argument of a verb(main) is shared with another verb(participle). The argument occurs only once in the sentence but is semantically related to both the verbs. The shared argument syntactically always attaches with the main verb. For the other verb this argument is semantically realized but not syntactically. Arguments of *-kara* constructions and *ke bAxa* constructions in Hindi-Urdu would fall under this type. Note the following sentence :

Non-adjectival-Shared-arg-DS-1 :

شاحد نے کھانا کھا کر پانی پیا

*SAhixa ne KAnA KAkara pAnI piyA*  
 Shahid Erg food having eaten water drank  
 ‘Shahid drank water after eating the food.’

It may be noted that linguistically *SAhixa ne* is explicit *karta* of only *piyA* ‘drank’ and not of *KAkara* ‘having eaten’, even though, semantically it is the agent for both *KAkara* and *piyA*. Since agreement and its vibhakti are controlled by the main verb ‘*piyA*’ (drank) it will be attached to it. However, its semantic presence of being an argument of ‘*Kakara*’ will be annotated by following the steps given above. After the annotation the inserted node would look as follows :

```
((      NULL__NP      <name='NULL__NP' dmrel='k1:'VGNF' reftype='corefn:NP'
mtype='non-gap'>
  NULL NN
))
```

SSF-13

'VGNF' and 'NP' in the values of attributes dmrel and reftype respectively are the names of the chunks to which this chunk would attach (VGNF) and would refer to (NP). Some more examples of this type of sharing are given below :

Non-adjectival-Shared-arg-DS-2 :

شاحد کھانا کھانے کے بعد پانی پیتا ہے

*SAhixa KAnA KAnE ke bAxa pAnI piWA hE*  
 Shahid food eating after water drinks be-Prs.Sg  
 ‘Shahid drinks water after eating food.’

Noun 'Shahid' in the above example is shared by '*KAnE*' (eating) and '*piWA\_hE*' (drinks)

The inserted chunk for 'SAhixa' in the above example would be :

```
((      NULL__NP      <name='NULL__NP' dmrel='k1:'VGNN' reftype='corefn:NP'
mtype='non-gap'>
  NULL NN
))
```

SSF-14

Non-adjectival-Shared-arg-DS-3 :

شاحد دلی جانا چاہتا ہے  
SAhixa xilll jAnA cAhawA hE  
Shahid delhi to-go want-hab be-Pres  
'Shahid wants to go to Delhi to Delhi.'

#### 4.3.2 Sharing in adjectival participles (wA\_huA constructions, KAye\_gaye constructions)

In another kind of sharing of arguments, a participle clause modifies the noun. and the modified noun, apart from being an argument of a higher verb, is also an argument of the verb in the participle clause. Therefore, the noun is shared by the main verb and its modifier verb. The adjectival participle, obviously, does not have the modified noun as its explicit argument. Again, although the argument in this case also is semantically realized but cannot occur syntactically. For example,

Adjectival-Shared-arg-DS-1 :

بچہ پر بیٹھا ہوا لڑکا سیب کھا رہا ہے  
bEnca para bETA huA ladZakA seba KA rahA hE  
bench on sit-perf be-ptpl boy apple eat prog pres  
'The boy sitting on the bench is eating an apple.'

\*Adjectival-Shared-arg-DS-2 :

میرے ذریعے لائے گے پھل اچھے تھے  
mere zariye lAye gaye Pala acCe We  
My-obl by buy-perf go-Perf fruits good past  
'The fruits bought by me were good.'

In example (Adjectival-Shared-arg-DS-1) above, bETA huA 'sit-perf be-ptpl' is modifying the noun ladZakA 'boy'. Noun ladzakA 'boy' is an argument of the higher verb KA rahA he 'eat prog pres'. ladZakA 'boy' is also an argument of the non-finite verb bETA huA 'sit-perf be-ptpl'. Similarly, in example (Adjectival-Shared-arg-DS-2) the noun Pala 'fruits' is an argument of both, the finite verb We 'were' and the non-finite verb lAye 'bought'.

As in the case of shared arguments of the non-adjectival participles, the arguments of this type will also be annotated. However, for such shared arguments, a new node will not be created. Instead, it will be captured by the label on the arc between the modifying clause and the modified noun. For example, the karaka relation of ladZakA 'boy' with KAwA huA 'eat.Impf.Ptpl' (in Adjectival-Shared-arg-DS-1) is k1 (karta karaka relation), it will be represented as nmod\_\_k1inv. Similarly, in example (Adjectival-Shared-arg-DS-2), KAye gaye 'ate go-Prf.' is the participle which modifies the noun Pala 'fruit', the noun Pala 'fruit' is k2 (karma karaka relation) of the verb Kaye hue 'eaten'. The relation between Pala 'fruits' and KAye hue 'eaten' will be represented as nmod\_\_k2inv.

Therefore, we have one more tag 'nmod\_\_k\*inv, which means nmod of the

type  $k^*inv$ , where  $k^*$  stands for the type of karaka relation i.e.  $k1$  or  $k2$  etc. and  $inv$  stands for inverse. Along with the karaka relation we also specify  $inv$  which denotes that, here the relation arc is going from child to the parent instead of parent to the child. In this type of sharing a new node is not created, the label  $nmod\_k^*inv$  is sufficient.

Adjectival-Shared-arg-DS-3 :

ڈالیوں پر کھلے پھول مہک رہے تھے  
*dAliyoM para Kile Pula mahaka rahe We*  
 branches on blossomed flowers smell prog past  
 'The flowers flowering on the branches were spreading  
 a scent.'

In the above example, Pula 'flowers' is the shared argument. Verb Kile 'blossomed' is modifying Pula 'flowers'. The feature structure of Kile 'blossomed' would be as follows :

```
(( VGNF <name='VGNF' drel='nmod__k1inv'>
  Kile VM
))
```

SSF-15

Since in this case, a new node is NOT inserted, none of the attributes which are annotated in an inserted node will be annotated here.

## 5. Some Additional Features

During the discussion on what all information would be useful for various applications, it was decided to add two more features on every finite verb clause. The two features are :

### 5.1 stype (Sentence type)

The attribute 'stype' is to be annotated on every finite verb chunk. The values for this are : declarative, imperative, interrogative etc. A complete list of the sentence type is provided separately. For example,

Sentence-type-DS-1:

آپ دروازہ بند کر دیں  
*Apa xaravAzA banxa kara xeM*  
 you door close do imp  
 'You close the door.'

The attribute 'stype' will be marked on the verb chunk. Thus, the annotated verb chunk with the 'stype' attribute would be as follows :

```
(( VGF <stype=imperative>
  kara VM
```

xeM VAUX

))

SSF-16

## 5.2 voicetype (Voice type)

The other feature to be annotated on every finite verb chunk is 'voicetype'. The values for this are only two (1) active and (2) passive. For example,

Voice-type-DS-1:

میٹنگ میں اس کا نام لیا گیا

mITinga meN usa kA nAma liyA gayA

meeting in his name take-perf go-perf

'His name was taken in the meeting.'

The voice type feature would be annotated on the verb as follows :

(( VGF <voicetype=passive>

liyA VM

gayA VAUX

))

SSF-17

Voice-type-DS-2:

اس نے دروازہ بند کیا

us ne xarvAza banxa kiyA

he door close do-p

'He closed the door.'

(( VGF <voicetype=active>

banxa VM

kiyA VAUX

))

SSF-18

## 5.3 coref (Corefer)

As mentioned in the section 4.1.28, relative clauses are attached to the noun they modify with a label 'nmod\_\_relc'. The attachment is between the main verb of the relative clause and the noun it modifies. Thus, an important information about the relative pronoun playing a crucial role in this relation is missed out. To capture this information, it has been decided to annotate the relative pronoun of the relative clause with an additional attribute of 'coref'. The value for the attribute 'coref' would be the referent noun in the main clause, i.e. the noun modified by the relative clause. An example of the same is :



Relative\_clause-DS-1:

میری بہن جو دلی میں رہتی ہے کل آ رہی ہے

merI bahana [ jo xillI meM rahawI hE] kala Aa rahI hE

my sister who Delhi in live-hab pres tomorrow come prog pres  
'My sister who lives in Delhi is coming tomorrow'

In the above example, the relative pronoun will, in addition to other features will also be marked with the attribute coref. Thus,

(( NP <name=NP>

merI

bahana

))

(( NP <coref=NP

jo

))

SSF-18

## 6. PART – 2 : Urdu Example Constructions

\*To be filled in with annotated examples.

اگر تلگو دیشم انتخابات میں ٹی آ ر ایس کے ساتھ تعاون نہ کرتی تو دس نشستوں پر بھی اس کو کامیابی نہ ملتی۔

If Telugu Desham had not coordinated with Congress then it could not have gained victory even in ten seats.

[[ [[ اگر \_CC ]] \_CCP [[ تلگو دیشم \_NNP انتخابات \_NN میں \_PSP ]] \_NP [[ ٹی \_NNPC آ \_NNPC ایس \_NNP  
کے \_PSP ساتھ \_NST ]] \_NP [[ تعاون \_NN ]] \_NP [[ نہ \_NEG کرتی \_VM ]] \_VGF [[ تو \_CC ]] \_CCP [[ دس \_QC  
نشستوں \_NN پر \_PSP بھی \_RP ]] \_NP [[ اس \_PRP کو \_PSP ]] \_NP [[ کامیابی \_NN ]] \_NP [[ نہ \_NEG ملتی \_VM -  
\_SYM ]] \_VGF ]] \_SSF

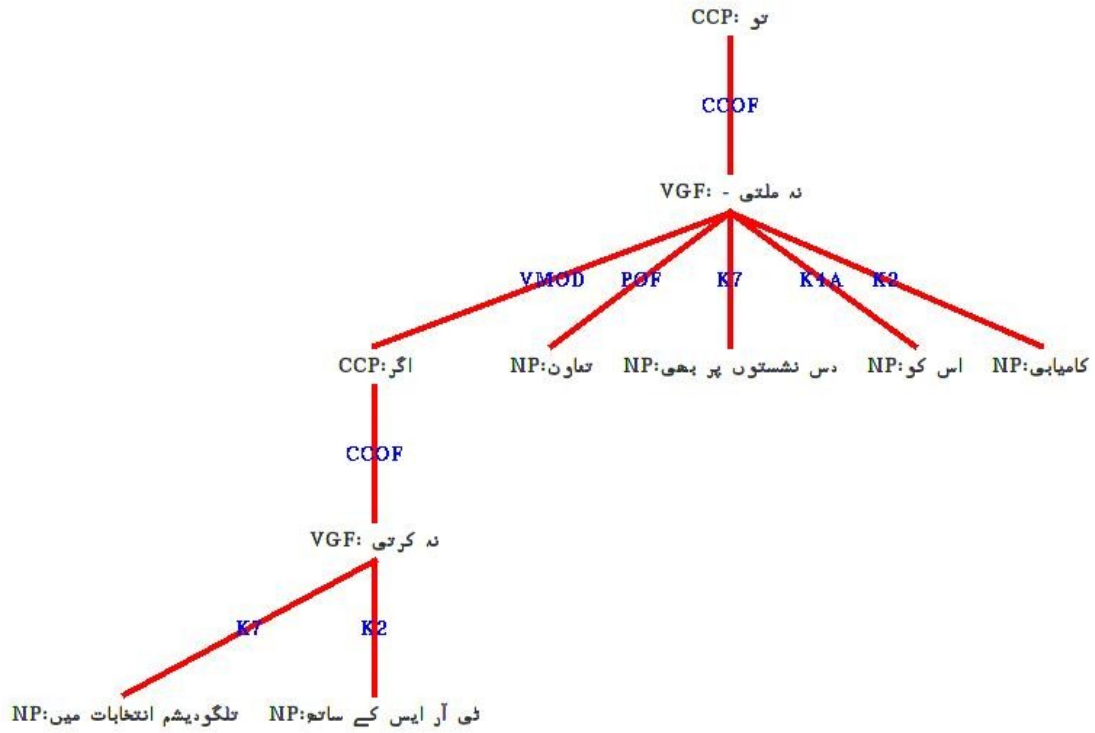


Figure 9

آج یہاں پارٹی ذرائع نے یہ بات بتائی۔

today here party sources-erg this matter reveal-past  
today, party sources have revealed this matter here.

[[ [[ آج\_NN ]\_NP [[ یہاں\_PRP ]\_NP [[ پارٹی\_NNC ذرائع\_NN نے\_PSP ]\_NP [[ یہ\_DEM بات\_NN ]\_NP  
[[ بتائی\_VM -\_SYM ]\_VGF ]\_SSF

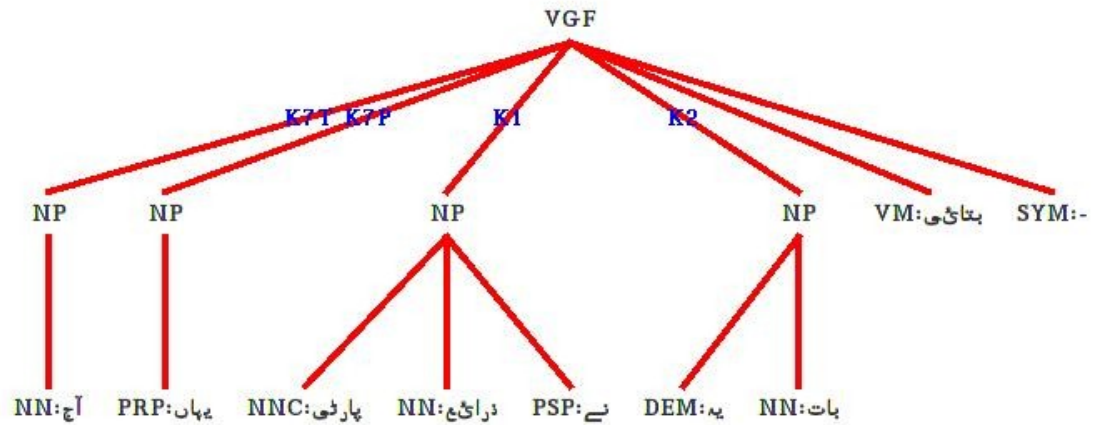


Figure 10

فوج نے ایک بیان میں یہ بات بتائی۔

military-erg one declaration in this matter reveal past  
the military has revealed this matter in a declaration.

[[ [[ آج \_NN ]]\_NP [[ یہاں \_PRP ]]\_NP [[ ذرائع \_NNC پارٹی ]]\_NP [[ نے \_PSP ]]\_NP [[ یہ \_DEM بات \_NN ]]\_NP  
[[ بتائی \_VM - \_SYM ]]\_VGF ]]\_SSF

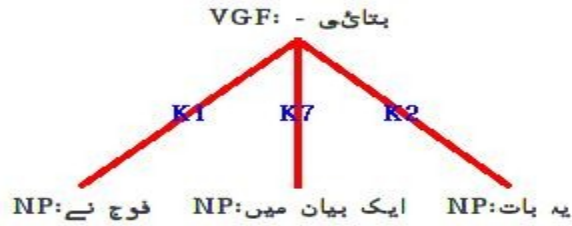


Figure 11

فوج کو اپنی پیشرفت سے قبل اس علاقہ میں بجھائے گئے دھماکو مادوں کو صاف کرنا پڑ رہا ہے۔

military -erg own proceeding from before this region

in lay past explosive substances Dat wipe do imperative.

The military has to wipe off the laid explosive bed before it takes measures.

[[ [[ فوج \_NN کو \_PSP ]]\_NP [[ اپنی \_PRP ]]\_NP [[ پیشرفت \_NN سے \_PSP قبل \_NST ]]\_NP [[ اس \_DEM  
علاقہ \_NN میں \_PSP ]]\_NP [[ بجھائے \_VM گئے \_VAUX ]]\_VGNF [[ دھماکو \_JJ مادوں \_NN کو \_PSP ]]\_NP [[  
صاف \_JJ ]]\_JJP [[ کرنا \_VM ]]\_VGNN [[ رہا \_VAUX ہے \_VAUX - \_SYM ]]\_VGF ]]\_SSF

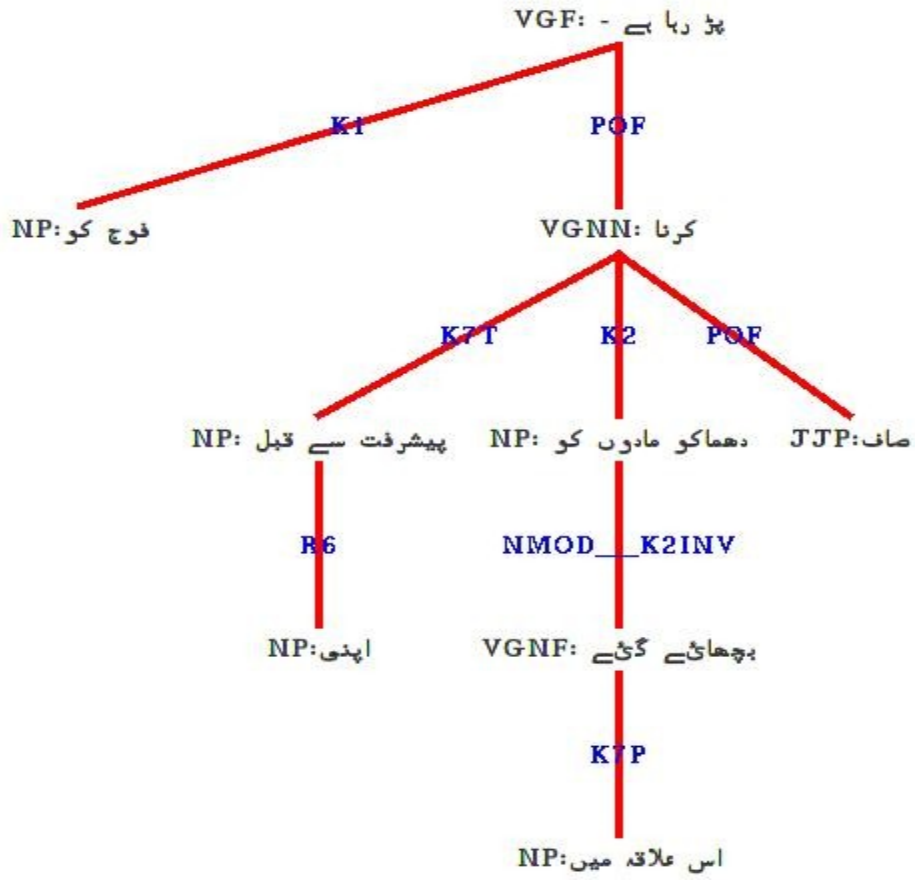


Figure 12

انہوں نے کہا کہ اس علاقہ میں 5 بڑے چمڑے کے کارخانے تھے۔ جو کنور موضع بھونگیر کو منتقل ہو چکے ہیں۔

they say past that this region in 5 large leather -gen industries be-pl past, which kanaur village bhongir Dat shift happen past participle.

They said that there were 5 large industries of leather which shifted to the kanaur Bhongir town.

[[ [[ انہوں \_PRP نے\_PSP ]]\_NP [[ کہا \_VM ]]\_VGF [[ کہ \_CC ]]\_CCP [[ اس \_DEM علاقہ \_NN  
 میں\_PSP ]]\_NP [[ 5\_QC بڑے \_JJ چمڑے \_NN کے\_PSP ]]\_NP [[ کارخانے \_NN ]]\_NP [[ تھے \_VM -  
 \_SYM ]]\_VGF [[ جو \_PRP ]]\_NP [[ کنور \_NN موضع \_NN بھونگیر \_NNP کو\_PSP ]]\_NP [[ منتقل \_JJ ]]\_JJF  
 ہو \_VM چکے \_VAUX ہیں \_VAUX - \_SYM ]]\_VGF ]]\_SSF

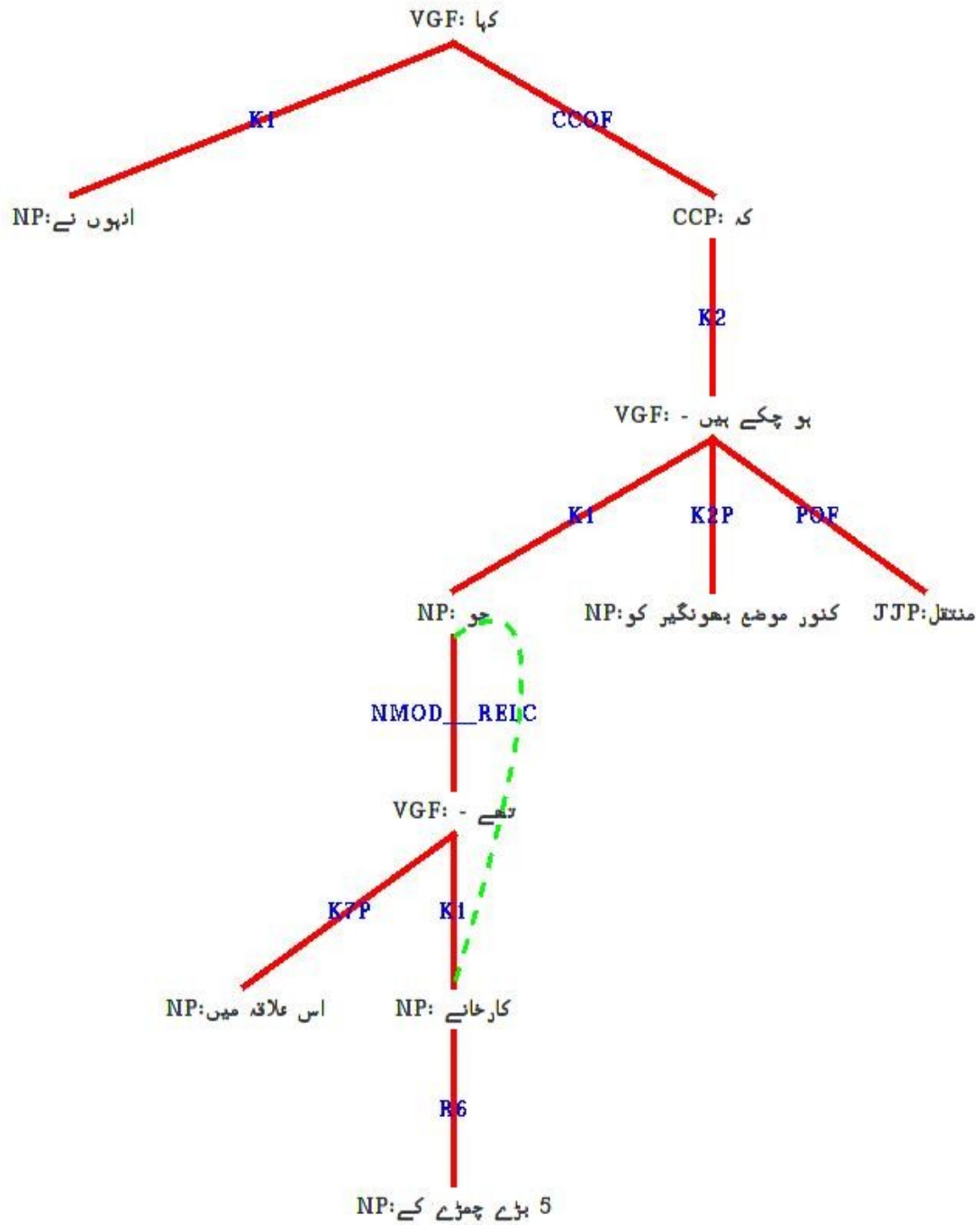


Figure 13

## 7. Conclusion

The tagging scheme presented above has been designed to annotate syntactic analysis within a dependency framework. The task of annotation for Hindi is underway and an initial experimental annotation on a part of the Urdu corpus has been completed. The Urdu annotators will now be working as per these guidelines. The basic scheme developed initially has been improved and revised.

## 8. References:

- R. Begum, S. Husain, A. Dhvaj, D. M. Sharma, L. Bai, and R. Sangal. 2008. Dependency annotation scheme for Indian languages. In Proceedings of IJCNLP-2008.
- A. Bharati, V. Chaitanya and R. Sangal. 1995. Natural Language Processing: A Paninian Perspective, Prentice-Hall of India, New Delhi, pp. 65-106.
- A. Bharati, D. M. Sharma, L. Bai and R. Sangal. 2006. AnnCorra : Annotating Corpora Guidelines For POS And Chunk Annotation For Indian Languages. LTRC Technical Report-31
- A. Bharati, R. Sangal and D. M. Sharma. 2007. SSF: Shakti Standard Format Guide. LTRC Technical Report-33
- Rajesh Bhatt. 2008. A Lecture at EFLU, Hyderabad.  
<http://people.umass.edu/bhatt/papers/eflu-aug18.pdf>
- M. Butt. 2004. The Light Verb Jungle. In G. Aygen, C. Bower & C. Quinn eds. Papers from the GSAS/Dudley House Workshop on Light Verbs. Cambridge, Harvard Working Papers in Linguistics, p. 1-50.
- D. Chakrabarty, V. Sarma and P. Bhattacharyya. 2007. Complex Predicates in Indian Language Wordnets, Lexical Resources and Evaluation Journal, 40 (3-4), 2007.
- E. Hajicova. 1998. Prague Dependency Treebank: From Analytic to Tectogrammatical Annotation. In Proc. TSD'98.
- M. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank, Computational Linguistics 1993.
- T. Mohanan, 1994. Arguments in Hindi. CSLI Publications.
- J. R. Ross. 1967. Constraints on variables in syntax, doctoral dissertation, MIT (published as 'Infinite syntax!' Ablex, Norwood (1986)).