Complex Predicates in Indian Language Wordnets

Pushpak Bhattacharyya Debasri Chakrabarti Vaijayanthi M. Sarma [1]

Abstract

Wordnets, which are repositories of lexical semantic knowledge containing semantically linked synsets and lexically linked words, are indispensable for work on computational linguistics and natural language processing. While building wordnets for Hindi and Marathi, two major Indo-European languages, we observed that the verb hierarchy in the Princeton Wordnet was rather shallow. We set to constructing a verb knowledge base for Hindi, which arranges the Hindi verbs in a hierarchy of is-a (hypernymy) relation. We realized that there are unique Indian language phenomena that bear upon the lexicalization vs. syntactically derived choice. One such example is the occurrence of conjunct and compound verbs (called C*omplex Predicates*) which are found in all Indian languages.

This paper presents our experience in the construction of lexical knowledge bases for Indian languages with special attention to Hindi. The question of storing or deriving complex predicates has been dealt with linguistically and computationally. We have constructed empirical tests to decide if a combination of two words, the second of which is a verb, is a complex predicate or not. Such tests will provide a principled way of deciding the status of complex predicates in Indian language wordnets. An additional application of this work is the possibility of automatic augmentations to the Wordnet using corpora, a topic of great interest in current research.

Keywords: Complex predicates, Wordnet, ontology, noun incorporation, compound verbs, automatic augmentation of wordnet, verb hierarchy

1. Introduction

Complex predicates (CPs) (also known as *complex verbs*) abound in South Asian languages (Hook [4], Mohanan [13], Verma [39] and Alsina [40]). They occur in the form of *nominal+verb* combinations (called *conjunct verbs*) and *verb+verb* combinations (called *compound verbs*). The key questions that we seek to resolve are:

A. *Given a N(oun)+V(erb) combination, is the noun incorporated into the verb complex or is it an overt argument of the verb?*
B. *Given a V(erb)+V(erb) combination, is the second verb an aspectual/modal or is it the polar (intensifier) in a polar-vector combination?*

These linguistically important and interesting questions are also highly significant for computational linguistics (CL) and natural language processing (NLP). Their computational significance arises from the issue of their storage in lexical resources such as wordnets (Fellbaum [1]) and ontologies (Guarino [2]) and raises the following questions:

C. *Given a corpus, how are complex predicates to be detected, thus paving the way for their automatic incorporation into the lexical knowledge network?*
D. *How exactly should complex predicates be stored, keeping in mind access and storage efficiency?*

[1] pb@cse.iitb.ac.in, Dept. of CS&E, debasri@cse.iitb.ac.in and vsarma@iitb.ac.in, Department of H&SS, IIT Bombay.

Deleted: Humanities and Social Sciences

This paper addresses these questions and presents our experience in the construction of lexical knowledge bases for Indian languages with special attention to Hindi. The question of storing or deriving complex predicates has been dealt with linguistically and computationally. We have constructed empirical tests to decide if a combination of two words, the second of which is a verb, is a complex predicate or not. Such tests will provide a principled way of deciding the status of complex predicates in Indian language wordnets. An additional application of this work is the possibility of automatic augmentations to the Wordnet using corpora, a topic of current research interest.

1.1 Hindi verbs

calnaa (*walk*), *caRhnaa*[2] (*climb*) *etc.* are examples of *simple verbs* in Hindi, and *aarambh karnaa* (*start*), *likh Daalnaa* (*write up*) *etc*. are examples of *complex verbs. Complex verbs* themselves are of two kinds: *conjunct verbs* (see 1 and 2 below) and *compound verbs* (see 3 and 4 below).

1. usne sabhaa **aarambh** **kii**
   (s)he-erg convention start do
   '(S)he started the convention.'
2. usne bacce ko **shaant** **kiyaa**
   (s)he-erg child acc peaceful do
   '(S)he pacified the baby.'
3. usne saaraa khaanaa **khaa** **liaa**
   (s)he-erg all food eat take
   '(S)he ate up all the food.'
4. vah ghar **calaa** **gayaa**
   (s)he home go past
   '(S)he went home.'

Example 1 contains a *noun+verb* sequence, while 2 contains an *adjective+verb* sequence. In examples 3 and 4, the first component of the verb complex is the main verb which occurs in its stem or other nonfinite form, while the second component is the 'helping' verb that carries the verbal inflections. Figure 1 shows the taxonomy of Hindi verbs [3].

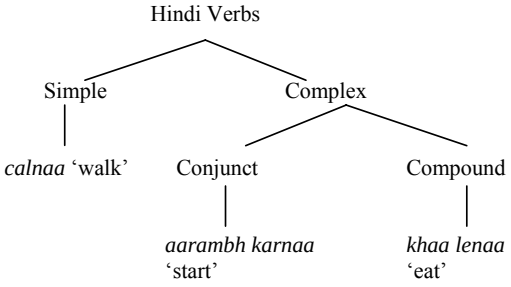

Figure 1: Taxonomy of Hindi verbs

In sections 1.2 and 1.3, we describe conjunct and compound verbs briefly. In section 2, we describe work related to this paper. Section 3 presents the background work on Hindi and

[2] Capital letters are used to represent the retroflexed series of consonants of Hindi

Marathi Wordnets (HWN and MWN) and the Hindi verb knowledge base (HVKB), which gave rise to the computational questions (C) and (D) above**.** Section 4 discusses complex predicates and the empirical tests designed to identify them. Section 5 concludes the paper.

1.2 N + V

Consider the word *help* in English, which appears in two different constructions:

5. He helped me with the housework.
6. He gave me help with the housework.

The equivalent in Hindi for *give help* is *madad (help) karnaa (to do)*. In Hindi, structures such as the one in 6 are possible while the one in 5 is not (i.e., direct noun to verb derivation). The question, therefore, is which word(s) is (are) to be stored in a lexical knowledge base (such as the Hindi Wordnet), *madad* or *madad karnaa* or both? There are two possible solutions:

a. Enter *madad* in the lexicon and then link it with the verb with which it co-occurs, i.e., *karna*
b. Enter *madad karnaa* as a single entry in the lexicon and then specify its syntactic and semantic features separate from either *madad* or *karna* individually specified

The first approach is the simplest. Syntactically there is no problem in terms of the argument structure of the associated verb and its subject-verb agreement properties. But, the meaning of the conjoined term is not strictly compositional. Consider, for example, *chalaang maarnaa 'dive'*. *maarnaa* may mean either *'to beat' or 'to kill'*. But neither meaning of the verb surfaces in *chalaang maarnaa.*

The obvious problem with the second solution is one of proliferating lexical items, redundantly; we cannot store every single N+V combination in the lexicon. Generally, the noun is a true object and there is no need to store it as a lexical unit along with a co-occurring verb. Thus it is necessary to separate true *conjunct verbs* from other similar looking *N+V* sequences. Consider the two combinations *caae (tea) lenaa (to take)* meaning '*to take tea'* and *jamhaaii (yawn) lenaa (to take)* meaning '*to yawn'*. In the former case *caae* (*tea*) is an overt object of the verb whereas in the latter, *jamhaaii* (*yawn*) is not.

1.3 V + V

V+V constructions are also difficult to deal with and describe, since there are many serial verb sequences to be found in the language. Consider,

7. raam kitaab **parh** **rahaa** **hai**
   Ram book read (V1) stay (V2) is
   'Ram is reading the book.'
8. raam-ne kitaab **paRh** **Daalii**
   Ram-erg book read (V1) pour (V2)
   'Ram (somehow) read (and finished) the book.'

rahaa in 7 marks the progressive aspect, whereas *Daalii* in 8 indicates the attitude (*somehow finished*). The V+V sequence in 7 should not be stored in the lexicon, whereas the one in 8 should probably be (for reasons that we discuss later). As in the case of conjunct verbs, it is necessary to separate true *compound verbs* from other *V+V* sequences.

2. Related work

2.1. Study of complex predicates

Constructions variously described as *complex predicates, compound verbs, composite predicates*, or *serial verbs* range across a number of expressions in diverse languages. Hook [4] summarizes his main conclusions about the function of the Hindi compound verb:

> "…(1) the compound verb expresses perfective aspect and is the marked member of a privative opposition perfective-imperfective (2) one of the functions of the compound verb is to express the completion of one action relative to the completion of another (3) if a verb is stative then it is non-compound (4) if there is no possibility of an action of event's being anticipated, it is expressed with a non-compound verb."

Hook [5] places the second verb, *i.e.*, the light verb, in an aspectual complex on par with the other auxiliaries in the language.

Butt [6] proposes some criteria to enable cross-linguistic identification of CPs:

> "The argument structure is complex (two or more semantic heads contribute to it).
> The grammatical function structure is that of a simple predicate (there is only a single subject and no embedding).
> The CP sequence may be formed either morphologically (lexically) or syntactically."

Butt [7] makes a distinction between *permissive, instructive* and *aspectual* complex predicates and takes the stand that complex predicate formation in Urdu takes place at *argument structure*.

Bashir [8] identifies the compound verbs of *Kalasha* and gives a semantic analysis based on 'prepared' and 'unprepared mind'. She proposes that the function of a compound verb is to grammaticize categories which involve the *knowledge* and *belief state* of the speaker.

Fedson [9] analyzes the semantic structure of the complex predicates in Tamil and classifies them into groups such as *stative, perfective, indication of status, aspectual etc.*

Kachru [10] remarks that 'In most of the South Asian languages, certain serial verbs express speaker attitudes, intentions, judgements, beliefs etc.' According to her, some of the serial verbs have a regular monoclausal or biclausal derivation and specific meanings such as disgust, disapproval, *etc.* She concludes that there is a need to distinguish serial verbs from other V+V sequences.

Pandharipande [11] in her article on 'Serial Verb Construction in Marathi' points out that V_1 and V_2 are paired on the basis of their semantic compatibility, which is subject to syntactic constraints. She uses tests to show the split properties of the serial verb construction in Marathi. According to her, some syntactic phenomena such as *passivization, participialization* and *verb agreement* refer to or operate on only V_2 whereas others such as *causativization* operate on both.

Paul [12] in her thesis gives an account of *Bangla* CVs within the framework of *Head-Driven Phrase Structure Grammar*. She proposes that the selection of a V_2 by a V_1 is largely determined at the level of semantics because the two verbs will unify if and only if they are semantically compatible.

Mohanan was the first to explicitly claim that Hindi has incorporation [13]. In doing so, she draws on morphosyntactic as well as semantic evidence. Dayal also notices this fact and concludes that incorporated nouns are syntactically visible and can function like regular arguments - at least for the purpose of agreement [14]. Mohanan too identifies this phenomenon and remarks, "In complex predicates, if the nominal predicate is logically transitive, and its

second participant is a logical object, this object is selected as the grammatical object of the complex predicate."[15]

2.2. Lexical databases

Wordnets are now considered indispensable resources for CL and NLP. The first wordnet created was for English at Princeton [16]. Eurowordnet, a linked multiwordnet, soon followed suit [17]. In the creation of lexical networks and ontologies, verbs have not received as much attention as they deserve. Ancient Sanskrit treatises on ontology like the *Amarkosha* [18] deal meticulously with nouns, but not with verbs. The present day ontologies and lexical knowledge bases such as *CYC* [19], *IEEE SUMO* [20], *WordNet* [1,16], *EuroWordNet* [17], *Hindi Wordnet* [21], *Framenet* [22] *etc.* build deep and elaborate hierarchies for nouns, but the verb hierarchies are either not present or, if present, are too shallow. The *Verbnet* project [23] is concerned exclusively with verbs and builds a very useful structure, but does not concern itself with building a *hierarchical structure.*

Classifying verbs and placing them in a structure according to their selectional preferences and other semantic properties are essential tasks in most text information processing tasks [23, 24] like machine translation, information extraction *etc*. Additionally, *property inheritance* (*e.g. walk* inherits the properties of *move*) facilitates lexical knowledge building in, for example, a rule based natural language analysis system [25].

Automatic augmentation to wordnets and ontologies are research areas of great interest currently [26, 27]. They, however, deal with named entities (proper nouns) and link them with the *instance of* relation ([27] also addresses automatic *meronymy* linking.) To our knowledge no work addresses automatic augmentation of verbs, and certainly not that of complex verbs.

3. Background work: Creation of lexical resources

We have, for long, been engaged in building lexical resources for Indian languages with focus on Hindi and Marathi (http://www.cfilt.iitb.ac.in). The Hindi and Marathi wordnets (HWN and MWN) [28] and the Hindi Verb Knowledge Base (HVKB) [29] have been given special attention. The wordnets more or less follow the design principles of the Princeton Wordnet for English while paying particular attention to language specific phenomena (such as *complex predicates*) whenever they arise.

3.1. Hindi and Marathi wordnets (HWN and MWN)

HWN and MWN have been created with the current statistics given in Table 1:

	Total number of Synsets	Total number of unique words
Hindi Wordnet	23,067	48,725
Marathi Wordnet	11,908	18,093

Table 1: Current status of Hindi and Marathi Wordnets

The status of the other Wordnets is given in Table 2 for comparison.

	Total number of synsets	Total number of unique words
WordNet (2.1)	117,597	155,327
GermaNet (2004)	53,312	76,563
Multi Word Net (1.39)	32,700	58,000

Table 2: Status of other Wordnets

We have incorporated a supporting ontology to whose nodes the synsets are linked and whose details are as follows:

Part of speech	Number of nodes
Noun	151
Verb	39
Adjective	35
Adverb	14

Table 3: Details of ontology

While the HWN has been created by manually looking up the various listed meanings of words in different dictionaries, the MWN has been created derivatively from HWN. That is, the synsets of HWN are adapted to MWN via addition or deletion of synonyms in the synset.

HWN entry:
{peR, vriksh, paadap, drum, taru, viTap, ruuksh, ruukh, adhrip, taruvar} '*tree*'
jaR,tanaa, shaakhaa, tathaa pattiyo se yukt bahuvarshiya vanaspati
'*perennial woody plant having root, stem, branches and leaves*'
peR manushya ke lie bahut hi upayogii hai '*trees are useful to men*'

MWN entry:
{jhaaR, vriksh, taruvar, drum, taruu, paadap} '*tree*'
mule, khoR, phaanghaa, pane ityaadiinii yokt asaa vanaspativishesh
'*perennial woody plant having root, stem, branches and leaves*'
tii damuun jhaadacyaa saavlit baslii '*Being tired/exhausted she sat under the shadow of the tree*'

Figure 2: MWN synset creation

Figure 2 shows the creation of the synset for the word *peR* 'tree' in MWN via addition and deletion of synonyms from HWN. The synset in HWN for this word is {*peR, vriksh, paadap, drum, taru, viTap, ruuksh, ruukh, adhrip, taruvar*} *'tree'*. MWN deletes {*peR,viTap, ruuksh, ruukh, adhrip*} and adds {*jhaaR*} to it. Thus, the synset for tree in MWN is {*jhaaR, vriksh, taruvar, drum, taruu, paadap*} *'tree'*. Hindi and Marathi being close members of the same language family, many Hindi words have the same meaning in Marathi. This is especially so for *tatsam* words, which are directly borrowed from Sanskrit. The semantic relations can be borrowed directly, thus saving time and effort.

3.1.1. Synsets

The principles of *minimality*, *coverage* and *replaceability* govern the creation of the synsets:
(i) Minimality: Only the minimal set that uniquely identifies the meaning is used to create the sysnet, *e.g.*,

{ghar, kamaraa} (*room*)

ghar- which is ambiguous- is not by itself sufficient to denote the concept of a *room.* The addition of *kamaraa* to the synset brings out this meaning uniquely.
(ii) Coverage: The synset should contain all the words denoting a particular meaning. The words are listed in order of (decreasing) frequency of their occurrence in the corpus.

{ghar, kamaraa, kaksh} (*room*)

(iii) Replaceability**:** The words forming the synset should be mutually replaceable in a specific context. Two synonyms may mutually replace each other in a context C, if the substitution of the one for the other in C does not alter the meaning of the sentence. Consider,

{svadesh, ghar} (*motherland*)– {apanaa desh} (*the country where one is born*)
amerikaa meN do saal bitaane ke baad shyaam svadesh/ghar lauTaa
America in two years stay after Shyam motherland returned
'*Shyam returned to his motherland after spending two years in America'*

The replaceability criterion is observed with respect to synonymy (semantic properties) and not with respect to the syntactic properties (such as subcategorization) of a lexeme. For instance, the two verbs {*aanaa, jaananaa*} *'know*' appear in the same synset for the word *know*. In Figure 3, the sentence frames show that while *aanaa* 'know' assigns dative case to the *subject NP*, *jaananaa* 'know' assigns nominative case. The two verbs {*aanaa, jaananaa*} *'know'* denote the same concept and each may replace the other in this particular semantic context.

HWN entry:
{aanaa, jaananaa} '*know*'
kisii kaarya ko karne me samarth honaa '*able to do something*'
Sentence Frames: NP1_DAT; NP2_NOM: mujhe silaii aati hai
1p,sg,DAT stiching know is
'*I know to stich*'
NP1_NOM; NP2_ACC: mai silaii jaanti huN
1p,sg,NOM stiching know is
'*I know to stich*'

Figure 3: Sentence frame for 'know'

A synset in HWN (and in MWN) consists of the following elements.

A. Synset: {vidyaalay, paaThshaalaa, skuul} (*school*)
B. Gloss: This consists of two parts.
   a. The text definition that explains the meaning denoted by the synset.
      vah sthaan jahaaM praathamik yaa maadhyamik star kii
      aupachaarik shikshaa dii jaatii hai
      'The place where formal education for primary or secondary level is given.'
   b. A sample sentence that uses the word in a sentence
      is vidyaalay meM pahalii se paanchavii tak kii shikshaa dii jaatii hai
      'Education from first to fifth class is given in this school.'

The data is stored in the Devanagari script in MYSQL database. The part of speech for each entry is listed in this database. Sample entries from both HWN and MWN are shown in Figure 4.

HWN entry:
{vidyaalay, paaThshaalaa, skuul} '*School*'
vah sthaan jahaan praathamik yaa maadhyamik star kii aupachaarik shikshaa dii jaatii hai
'*The place where the formal education of primary or secondary level is given*'
"is vidyaalay meM pahalii se paanchavii tak kii shikshaa dii jaatii hai"
'*Education from first to fifth class is given in this school*'

MWN entry:
{shaaLaa, vidyaalay, paaThshaaLaa} '*school*'
jethe praathamik va maadhyamik staraavarii aupachaarik shikshaN dile jaate te ThikaaN

Figure 4: HWN and MWN Sample Entry

3.1.2. Lexical relations

HWN incorporates several commonly used semantic and lexical relationships along with a few new ones. A brief description is given below:

a) Antonymy is a lexical relation indicating 'opposites'. For instance, {**moTaa**, sthuulkaay}'*fat*' → {**patlaa**, dublaa} '*thin*'

patlaa (*thin*) is the antonym of moTaa (*fat*) and vice versa. The HWN also indicates the criterion under which the antonomy holds. In the above example, the antonymy criterion is *size*. Other criteria are given in Table 4.

Criterion	Examples	Gloss
Size	(chhoTaa-badzaa, moTaa -patlaa)	big-small, thick-thin
Quality	(achchhaa-buraa, pyaar-ghriNaa)	good-bad, love-hatred
Gender	(beta-beTii, maataa-pitaa)	son-daughter, father-mother
State	(shuruu-ant)	beginning-end
Personality	(raam-raavaN)	Rama-Ravana
Direction	(puurv-pashchim, aage-piichhe)	eat-west, front-behind
Action	(lenaa-denaa, khariid-bikrii)	take- give, buy-sell
Amount	(kam-jyaadaa, halkaa-bhaarii)	little-much, light-heavy
Place	(duur-paas)	far-near
Time	(din-raat, subaha-shaam)	Day-night,morning-evening

Table 4: Criteria for Antonymy

b) Gradation is a lexical relation that represents possible intermediate states between two antonyms. Figure 5 shows the gradation relation among time words.

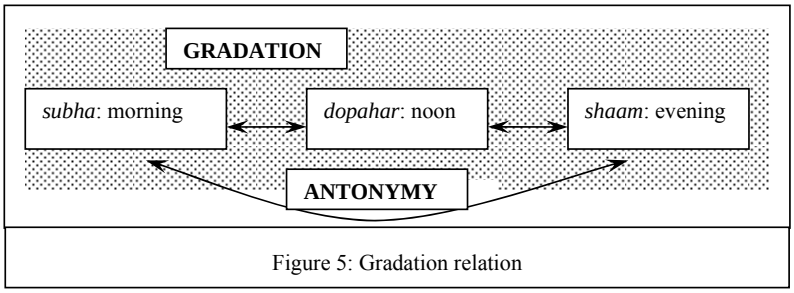

Figure 5: Gradation relation

c) Hypernymy and Hyponymy encode lexical relations between a more general term and specific instances of it.

{belpatra, belpattii, bilvapatra} '*leaf of a tree named bela*'
→ {pattaa, paat, parN, patra, dal} '*leaf*'

Here, *belpatra* (*a leaf of the bel tree*) is a kind of *pattaa* (*leaf*). *pattaa* is the hypernym of *belpatra* and *belpatra* is a hyponym of *pattaa*.

d) Meronymy and Holonymy express the *part-of relationship* and its inverse.

{*jaR, muul, sor*} '*root*' → {*peR, vriksh, paadap, drum*} '*tree*'

Here, *jaR* (*root*) is a part of *peR* (*tree*) and therefore, *jaR* is the meronym of *peR* and *peR* (*tree*) is the holonym of *jaR* (*root*).

e) Entailment is a semantic relationship between two verbs. Any verb *A* entails a verb *B*, if the meaning of *B* follows logically and is strictly included in the meaning of *A*. This relation is unidirectional. For instance, *snoring* entails *sleeping*, but *sleeping* does not entail *snoring*.

{*kharraaTaa lenaa, naak bajaanaa*} '*snore*'→ {*sonaa*} '*sleep*'

f) Troponymy is a semantic relation between two verbs when one is a specific 'manner' elaboration of another. For instance,

{dahaaRanaa} '*to roar*' is the troponym of {bolanaa} '*to speak*'

g) The HWN also cross-links synsets across different parts of speech. These links have not been taken from the EWN. Cross-links between 'nouns' and 'verbs' include the following:

i. Ability link specifies the features inherited by a nominal. For example,
{*machlii, macchii, matsya, miin, maahii*} '*fish*'→ {*tairnaa, pairnaa, paunrnaa*} '*swim*'

ii. Capability link specifies the features that may be acquired by a nominal. For example,
{*vyakti, maanas*} '*person*' → {*tairnaa, pairnaa, paunrnaa*} '*swim*'

iii. Function link specifies function(s) associated with a nominal. For example,
{*adhyaapak, shikshak*}'*teacher*' → {*paRhanaa, shikshaa denaa*} '*teach*'

Cross-links between 'nouns' and 'adjectives' are used to indicate typical properties of a noun. For example, {*sher*} '*tiger*' → {*maansaahaarii*} '*carnivorous*'. Links between morphologically derived forms mark the root form from which a particular word is derived by affixation. For example, {*bhaaratiiyataa*} '*indianness*' is derived from {*bhaaratiiya*} '*Indian*' and is linked to it. Figures 6 and 7 below show the web interfaces for HWN and MWN.

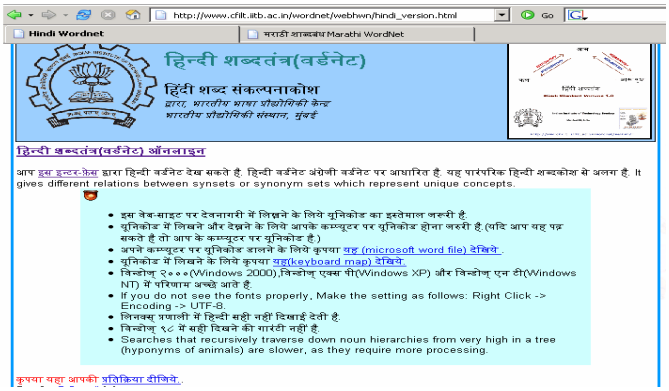

Figure 6: Web interface for Hindi Wordnet

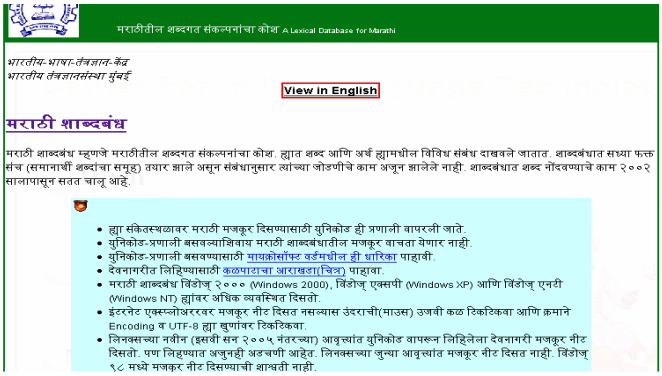

Figure 7: Web interface for Marathi Wordnet

3.1.3. Towards a multilingual Indo-Wordnet

While wordnets have been created for many different languages, they all ultimately do one thing: *they represent meanings via synsets and link them through semantic and lexical relations*. These meanings are mostly universal, the exceptions being culture-specific synsets. This fact suggests a way of automatically linking the synsets of a wordnet using another wordnet which is already complete in all respects − synset repository, semantic relations and lexical relations.

The process of setting up lexico-semantic relations in one wordnet using the corresponding information from another wordnet is called *Relation Borrowing*. The algorithm for relation borrowing is given in Figure 9 and there are three possible outcomes when deriving the MWN from the HWN. (a) When the meaning is found in both Hindi and Marathi, the relations are established in MWN for that meaning (Figure 9). This is the most common outcome, since Hindi and Marathi are sister languages (Indo-Aryan family) and exist in almost identical cultural settings. (b) When the meaning is found in Hindi but not in Marathi, relation borrowing is not possible. For instance, {*daadaa, baabaa, aajaa, daddaa, pitaamaha, prapitaa*} are words in Hindi for *paternal grandfather*. There are no equivalents for this meaning in Marathi. (c) When the meaning is not found in Hindi but is found in Marathi, the relations must be set up manually. For example, {*gudhipaadvaa*, *varshpratipadaa*} meaning *'new year'* are words in Marathi which do not have any equivalents in Hindi.

```
for each synset identity marathi_synset_id in Marathi WordNet do
   if (marathi_synset_id == hindi_synset_id) do
  for each relation r pointed by hindi_synset_id do
   if (relation type of r is semantic) do clamp the synset
      identity linked by relation r in to marathi_synset_id
   end if
   else
    clamp the synset identity linked by relation r in
    Hindi_synset_id to Marathi_synset_id AND manually insert
    the corresponding lexical element
   end else
  end for
 end if
```

Figure 8: Algorithm for relation borrowing between HWN and MWN

HWN and MWN implementations contain the following data structures:

a. A table called *tbl_all_words* which stores for each word the part of speech and an array of ids for the synsets in which the word participates. Tables 5 and 6 illustrate this for word *kar 'do'*-

Table 5: HWN tbl_all_words

hindi_synset_id	Word	PoS
491	*kar*	noun
3295	*kar*	verb
3529	*kar*	noun
4107	*kar*	noun
13314	*kar*	noun
13322	*kar*	noun
11958	*kar*	verb
11959	*kar*	verb
11960	*kar*	verb
11961	*kar*	verb
11962	*kar*	verb

Table 6: MWN

marathi_synset_id	Word	PoS
4107	*kar*	noun
4115	*kar*	verb

tbl_all_words

b. A table called *tbl_all_synsets* (Tables 7 and 8) which stores the sysnet ids, the synsets and the glosses of the various meanings.

hindi_synset_id	Synset	Gloss	Category
491	<not shown	<not shown	noun
3295	due to space	due to space	verb
3529	constraint>	constraint>	noun
4107			noun
13341			noun
13322			noun
11958			verb
11959			verb
11960			verb
11961			verb
11962			verb

Table 7: HWN tbl_all_synsets

marathi_synset_id	Synset	Gloss	Category
4107	<not shown	<not shown	noun
4115	due to space	due to space	verb
	constraint>	constraint>	

Table 8: MWN tbl_all_synsets

c. A table *tbl_<PoS>_<Relation>* for each PoS and Relation combination. For example, *tbl_noun_hypernymy* is the table for the semantic relation of hypernymy. Continuing the example for *kar*, Table 9 shows its hypernyms.

synset_id	*hypernymy_id*
491	503
3529	985
4107	3051
13341	12149
13322	1070
11958	2015
11959	3666
11960	7120

Table 9: HWN tbl_noun_hypernymy

Using the basic ideas outlined above, the synsets of MWN are completely linked with semantic and lexical relations. This saves a lot of manual labour. An interface has been designed to facilitate the simultaneous browsing of HWN and MWN. The input to this browser is a search string in any of the two languages. The search results for both the languages are displayed simultaneously.

3.2. Verb knowledge base (VKB)

As a part of our language processing resources, we are also building VKBs for both English and Hindi. The VKB contains a hierarchical arrangement of verbs. For the English VKB, we have extracted the verbs from the *British National Corpus (BNC)* [30] and for the Hindi VKB, we have extracted the verbs from the corpus of the Central Institute of Indian Languages, Mysore [31]. The different meanings of a verb are listed after consulting WordNet 2.1 [32], Oxford Genie [33], Hindi dictionaries [34, 35, 36, 37] and the corpus. The selectional restrictions and case properties are specified in Universal Networking Language (UNL) [38].

UNL is an electronic language for computers to express and exchange information. The UNL system is composed of *universal words (UWs)* (explained below), *relations, attributes* and the *UNL knowledge base (KB)*. The UWs constitute the vocabulary of the UNL, the *relations* and the *attributes* constitute the syntax and the UNL KB constitutes the semantics. The KB defines possible relationships between UWs. The UNL represents information for any sentence as a hypergraph with words forming the nodes and the arcs indicating the relations. The relations between UWs in binary relations have different labels depending on the semantic roles of each UW. Each relation label is a string of 3 characters or less. Any node in this structure may itself be a graph, in which case the node is called a *compound word (CW)*. Figure 10 shows the UNL representation for the sentence *John eats rice with a spoon*. In this figure, the relation arcs are labeled *agt* (agent), *obj* (object) and *ins* (instrument).

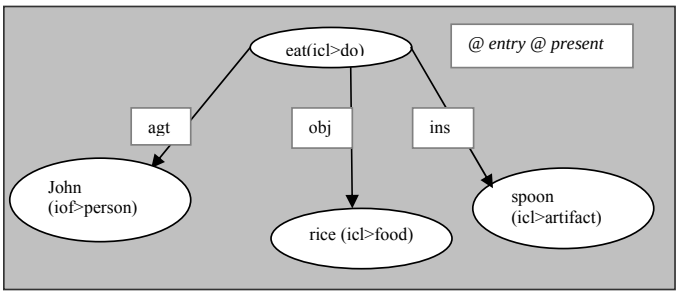

Figure 9: UNL graph of ***John eats rice with a spoon***

In its current manifestation the UNL contains 41 relation labels. The nodes *eat(icl>do), John(iof>person), rice(icl>food)* and *spoon(icl>artifact)* are the UWs. These are language words with *restrictions* given in parentheses. *icl* stands for *inclusion* and *iof* stands for *instance of*. UWs can be annotated with attributes such as *number*, *tense etc.* which provide further information about how the word is being used in that specific sentence. Any of the three restriction labels- *icl, iof* and *equ*- can be attached to an UW to restrict its sense.

After specifying the UNL relations, the sentence frames and semantic attributes of a verb are given. Figure 11 contains an example from the HVKB. We first list the verb, followed by the UNL relations that it enters into, an example sentence, gloss, the syntactic frame where the verb occurs and finally its grammatical and semantic attributes (VTRANS, VOA-ACT etc.). Currently there are approximately 2000 nodes in the HVKB and about 7000 nodes in the EVKB.

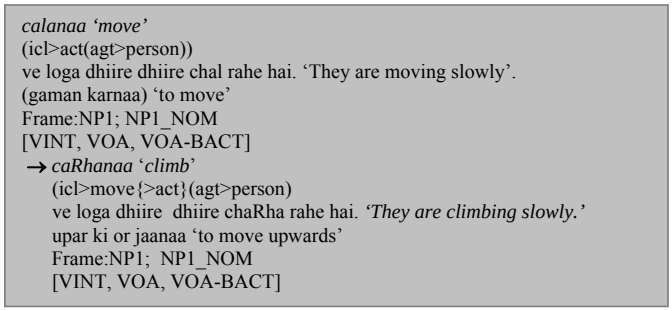

Figure 10: Partial hierarchy of 'move' class in HVKB

3.2.1 HVKB and HWN

The HVKB is built separately and exclusively for verbs. These verbs are then to be incorporated into the HWN. The HWN will use the subcategorisation frames from the HVKB and a verb's synsets will incorporate the elaborated grammatical and semantic information that is stored in the HVKB. Following the convention adopted in the EWN, we provide information about any alternative syntactic frames, should they arise, for the other members of a synset. This is shown in Figure 12. The two verbs {aanaa, jaananaa} '*know*' appear in the same synset for the word *know* but in different syntactic frames. The verb and its frame are marked with an asterisk to show the link between them.

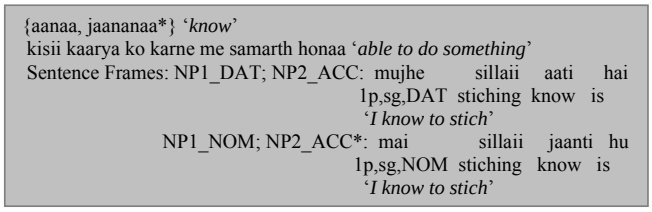

Figure 11: Sentence Frames in HWN

4. Complex predicates (CP)

As discussed in section 1, there are large numbers of multi-word verb constructions in Hindi. It is necessary to study these constructions carefully in order to store them systematically. In this section we will deal with N+V and V+V constructions.

4.1 Noun incorporation in Hindi verbs

Hindi has many *conjunct verbs* that need to be distinguished from regular object noun and verb sequences. In what follows we describe the empirical tests we use to distinguish between the two constructions.

A. Addition of the accusative case marker to the noun
B. Constituency tests, including Movement, conjunct question and coordination tests
C. Addition of modifiers to the noun phrase

4.1.1 Adding the accusative case marker

The aim of this test is to see whether the sentence is acceptable *to a native speaker of Hindi* after overtly marking the object with the accusative case-marker. While a true object noun will allow accusative marking, an incorporated noun will not. Consider the following examples,

9. a. raam ne caae lii
ram erg tea take-past
'Ram took tea.'

b. raam ne us caae ko liyaa jo khulii thii
ram erg that tea acc take which open was
'Ram took that tea which was kept open.'

10. a. raam ne jamhaaii lii
ram erg yawn take-past
'Ram yawned.'

b. * raam ne us jamhaaii ko liyaa …
ram erg that yawn acc take-past …
'Ram took that yawn.'

11. a. vah mere kaam me rucii letii hai
3p-sg-nom my work in interest take is
'She takes an interest in my work.'

b. *usne mere kaam me us rucii ko liyaa jo ..
3p-sg-erg my work in that interest acc take-past which....
'(S)he takes that interest in my work which ....'

In 9 the direct object of the verb *lenaa* (*take*) is *caae* (*tea*). In 9a the noun occurs without a case marker and in 9b it is overtly marked; both sentences are acceptable. In 10 and 11 the nouns *jamhaaii* (*yawn*) and *ruci* (*interest*) appear in the direct object position. In 10a and 11a they occur without the case marker; the sentences are acceptable. But in 10b and 11b they are overtly marked and the sentences are unacceptable. These nouns must be incorporated and not true syntactic objects.

4.1.2 Constituency tests

Three tests, *movement, conjunct response* and *coordination* are used to analyze the internal structure of conjunct verbs. .

4.1.2.1 Movement of the Noun

Object nouns in Hindi are usually free to occur in non-canonical positions. If the N+V combination resists such relocation then it must form a single constituent:

12. a. usne subaha uthkar **caae** lii
3p-sg-erg morning wake **tea** take-past.
'After waking up in the morning he took tea.'

b. **caae** usne subaha uthkar lii

13. a. kahaani-ne dukhaant **ruup** liaa
story-erg tragic shape take-past.
'The story took a form of tragic ending.'

b. ***ruup** kahaanii ne dukhaant liaa

14. a. usne pratiyogita meN **bhaag** liaa
3p-sg-erg competition in part take-past.
'(S)he took part in the competition'.

b. ***bhaag** usne pratiyogita meN liaa

In all the *b* examples, the noun has been moved from its canonical position. In 12b, *caae* (*tea*) is the actual object of the verb *lenaa* (*take*). Both 12a and b are acceptable. 13a and 14a are deemed grammatical while their *b* counterparts are not. The nouns in these cases must be incorporated into the verb and are not independent. *ruup lenaa* (*take shape*) and *bhaag lenaa* (*take part*) are therefore taken to be instances of conjunct verbs.

4.1.2.2 Constituent Response Test

Consider the following *N+V* sequences *jamhaaii lenaa* (literally, *yawn take* meaning *yawn*) and *chalaang maarnaa* (*jump beat* meaning *jump*).

15. raam ne jamhaaii lii
    ram erg yawn take-past
    'Ram yawned.'
16. raam ne chalaang marii
    ram erg jump beat-past
    'Ram dived.'

Questions on the actions in 15 and 16 reveal another property of incorporated nouns.

17. raam ne kyaa kiyaa
    What did Ram do? (NOT raamne kyaa liyaa, *What did Ram take*?)
    Answer: He yawned.
18. raamne kyaa kiyaa
    What did Ram do? (NOT raamne kyaa maaraa, *What did Ram beat*?)
    Answer: He jumped.

Compare these with 19 and 20:

19. vah bazaar se phal laayaa hai
    3p-sg-nom market from fruit brought is
    'He brought fruits from the market.'
20. a. vah kya laayaa *(What did he bring?)*
    b. usne kyaa kiyaa (*What did he do*?)

Unlike the questions in 17 and 18, the questions based on the action in 19 may isolate either the noun (object) (20a) or the whole verb phrase (20b). Thus the nouns in 17 and 18 have a different status from the one in 19.

4.1.2.3 Coordination Test

Coordination is possible only between identical syntactic constituents. This test shows that incorporated nouns resist coordination. In 21, *namkiin* (snack) can be conjoined with *caae* (*tea*). In 22 *niind* (*sleep*) cannot be conjoined with *jamhaai* (*yawn*) and in 23, *ksamaa* (*forgiveness*) cannot be conjoined with *vidaa* (*farewell*). However it is possible to conjoin the entire N+V combination as is seen in 23b. This test shows that the incorporated noun and the verb combination have the status of a single lexical or syntactic unit.

21. log [[caae] aur [namkin]] le rahe the
    people tea and snack take prog past
    'People were taking tea and snacks.'

22. *log [[niind] aur [jamhaaii]] le rahe the
people sleep and yawn take prog past
'People were taking sleep and yawn.'

23. a. *usne bacce ko [[ksamaa] aur vidaa]] kii
3p-sg-erg children acc forgiveness and farewell do
'(S)he forgave and bade farewell to the children.'

b. usne bacce ko [[ksamaa kiyaa] aur [vidaa kii]]
3p-sg-erg children acc forgiveness do and farewell do

4.1.3 Adding modifiers to the NP

True arguments may be modified by adjectives, determiners, numerals, sentences *etc*. but incorporated nouns do not permit such modifiers.

24. maine aaj **bahut** caae pii
1p-sg-erg today a-lot-of tea drink-past
'Today I had a lot of tea.'

25. usne **jor se** dhakka maaraa
3p-sg-erg adv push beat-past
'He pushed hard.'

26. usne merii **bahut** madad kii-past
3p-sg-erg me a-lot-of help do
'He helped me a lot.'

In 24 the modifier *bahut* (*a lot*) modifies the noun *tea* and not the whole verbal group but in 25, the modifier *jor se* (*heavily*) modifies *dhakkaa maarnaa* (*push-beat*) as a whole. Similarly, in 26 *bahut* (*a lot*) modifies *madad karnaa* (*help-do*) and not just either *madad (help)* or *karnaa* (*do*).

4.2 Compound verbs

We have identified 5 kinds of *V+V* sequences in Hindi. Not all of these are examples of complex predicates.

1. ***V_1 inf-e+ V_2*:** Here V_1 is in the infinitival form. V_2 is always the verb *lagnaa* (*literally, attach*) and it carries the inflections of *number, gender, tense etc*: *e.g.*, *rone lagnaa* (literally, *cry-attach* meaning *start crying*), *likhne lagnaa* (*write-attach* meaning *start writing*). The second verb yields the meaning of beginning of an action or happening of an event. Thus, *lagnaa* (*attach*) is treated as a modal auxiliary.

2. ***V_1 inf+ V_2*:** V_1 is in the infinitival form. V_2 is always the verb *paRnaa* (literally, *fall*). For instance, *bolnaa paRaa* (literally *say-fall* meaning *compelled to say*), *likhnaa paRaa* (*write-fall* meaning *compelled to write*). The second verb always gives the sense of force or compulsion. This is also treated as a combination of a verb and a modal auxiliary.

3. ***V_1 inf-pp+ V_2 stem*:** V_1 is in the infinitival form with the postposition (*pp*) *ke lie* (*for*). The *pp* may be shortened to *ko* (*to*). Both the forms *ko* and *ke lie* are used. For example, *likhne ko/ke lie kahaa* (*asked to write*), *paRhne ko/ke lie diyaa* (*gave to read*). In these constructions, the two verbs behave independently. This is seen when we apply different syntactic tests. For instance,

Scope of Adverb

27. a. us-ne mujhe khat **jaldi-se** likhne-ko kahaa
3p-sg-erg 1p-sg-Dat letter quickly write-inf -pp say-past
'He asked me to write the letter quickly.'

b. us-ne mujhe khat likhne-ko **jaldi-se** kahaa.
3p-sg-erg 1p-sg-Dat letter write-inf -pp quickly say-past
'He quickly asked me to write the letter.'

Scope of Negation

28. a. us-ne mujhe khat likhne-ko **nahiiN** **kahaa**
3p-sg-erg 1p-sg-Dat letter write-inf-pp not say-past
'He hasn't asked me to write a letter.'

b. us-ne mujhe khat **nahiiN** likhne-ko kahaa.
3p-sg-erg 1p-sg-Dat letter not write-inf-pp say-past
'He asked me not to write a letter.'

In 27a the adverb *jaldi se* (*quickly*) modifies the first verb *likhne* (*to write*) while in 27b, the scope of the adverb is over the second verb alone. In 28a the scope of negation is over the first verb and in 28b it is over the second verb. Thus, this type of *V+V* sequence does not behave as if it were a single, unalterable constituent.

4. ***V_1 –kar+ V_2*:** In this type of multi-verb construction the suffix *–kar* is attached to V_1. For example *likhkar ayaa* (*wrote and came*), *lekar gayaa* (*took and went*) etc. Here too, the verbs indicate two different actions.

Scope of Adverb

29. a. vah **jaldi se** nahaa-kar aaye-gaa
3p-sg-nom quickly bath-do come-fut.
'He will quickly take a bath and come.'

b. vah nahaa-kar **jaldi se** aaye-gaa.
3p-sg-nom bath-do quickly come-fut.
'He will take a bath and come quickly.'

Scope of negation

30. a. vah **nahiiN** nahaakar aye-gaa
3p-sg-nom not bath-do come-fut
'He will come without taking a bath.'

b. vah nahaakar **nahiiN** aye-gaa.
3p-sg-nom bath not come-fut
'He will not come after taking a bath.'

In 29a the scope of the adverb is over the first verb and in 29b the scope is over the second verb. In 30a the negative marker precedes the first verb and negates it and in 30b it precedes the second verb and negates it. Thus, this type of *V+V* sequence does not behave as a single constituent.

5. ***V_1 stem+ V_2*:** V_1 occurs in the bare form and V_2 bears all the inflections. Examples of such type are *maar Daalnaa* (literally *kill-put,* meaning *kill*), *likh lenaa* (literally *write-take,* meaning *write*). V_2 loses its primary meaning, but nonetheless adds some meaning to the whole sequence. Adverbial scope and negation tests show that it is not possible to treat these verbs individually; thus the scope of the adverb is over the whole sequence, as is the scope of the negation.

4.2.1 Compound verb (CV) sequence as a constituent

Following Butt [6] and Paul [12] we use the following tests to identify compound verbs. The tests are similar to the ones we outlined for conjunct verbs in section 4.1.

1. Scope of adverbs
2. Scope of negation

3. Nominalization
4. Passivization
5. Causativization
6. Movement

We have applied these tests to various data and elicited native speakers' judgements on the grammaticality of different sentences (as in section 4.1 for NI). Without going into the details, we present in Table 10 a summary of the results. These tests provide a reliable way in which to detect true CVs in the corpus and to differentiate them from other *V+V* combinations.

Verb Group	Test	Result
V1 stem+ V2	Scope of Adverb	over V1
V1 inf-e+ V2	---	over V1
V1inf+V2 stem	---	over V1
V1inf-pp+V2 stem	---	V1 or V2 depends upon the adverb's syntactic position
V1 –kar+ V2	---	V1 or V2 depends upon the adverb's syntactic position
V1 stem+ V2	Negation	No negation
V1 inf-e+ V2	---	Negates the whole sequence
V1inf+V2 stem	---	V1 or V2 depends upon the negative's syntactic position
V1inf-pp+V2 stem	---	V1 or V2 depends upon the negative's syntactic position
V1 –kar+ V2	----	V1 or V2 depends upon the negative's syntactic position
V1 stem+ V2	Nominalization	V2 bears nominal morphology, V1+V2 is nominalized
V1 inf-e+ V2	---	No nominalization
V1inf+V2 stem	----	No nominalization
V1inf-pp+V2 stem	----	No nominalization
V1 –kar+ V2	----	Both the verbs can be nominalized separately
V1 stem+ V2	Passivization	V2 bears passive morphology, V1+V2 is passivized
V1 inf-e+ V2	---	No passivization
V1inf+V2 stem	----	No passivization
V1inf-pp+V2 stem	----	V2 bears passive morphology, and only V2 is passivized
V1 –kar+ V2	----	Both the verbs can be passivized separately
V1 stem+ V2	Causativization	V1 bears causative morphology, V1+V2 is causativized
V1 inf-e+ V2	---	V1 bears causative morphology, V1+V2 is causativized
V1inf+V2 stem	----	V1 bears causative morphology, V1+V2 is causativized
V1inf-pp+V2 stem	----	Both the verbs can be causativized individually
V1 –kar+ V2	----	Both the verbs can be causativized individually
V1 stem+ V2	Movement	No movement
V1 inf-e+ V2	---	Stylistic movement
V1inf+V2 stem	----	No movement
V1inf-pp+V2 stem	----	Stylistic movement
V1 –kar+ V2	----	Stylistic movement

Table 10: Summary of results on CV identifications tests

4.3 CPs, HVKB and HWN

In the previous sections we have discussed different multiword verb constructions in Hindi which we refer to as CPs. We have shown how, with the use different syntactic and semantic tests, we isolate CPs from similar constructions. The semantic analysis of both conjunct and compound verbs show us that these sequences are the result of lexical derivational processes and, consequently, these items are to be directly stored in the lexical databases like HVKB and HWN. HVKB already stores these verb sequences and they will also be included in the HWN as the latter expands, provided these multiword verbs meet the criteria for CP-hood. Thus, in the HWN,

we will store both *nahaanaa 'bath'* and *snaan karnaa 'to take a bath'* in the same synset for the word *'bath'*.

5. Conclusion

We have described in this paper our work on complex predicates with a view to obtaining a principled way of deciding when and whether or not to include them in lexical knowledge bases. We offer systematic tests to identify complex predicates. The central question of when an *N+V* sequence or a *V+V* sequence become direct candidates for storage in the lexical knowledge base may be resolved by the application of these tests. Computational implementation of these tests should enable automatic augmentation of wordnets, a very desirable outcome. Our future work concerns such implementation of these empirical tests and we aim to integrate it with corpora search.

References

[1] Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*, MIT Press.

[2] Guarino, N. 1995. "Formal; Ontology, Conceptual Analysis and Knowledge Representation", International Journal of Human and Computer Studies, 43(5/6), pp. 625-640.

[3] Bahari H. 1997. *Vyavaharik Hindi Vyakaran Tatha Rachna,* Lokbharti Prakashan, Allahabad, India.

[4] Hook, P. E. 1974. "The Hindi compound verb: What it is and what it does?" In *Readings in Hindi-Urdu linguistics*, K.S. Singh (ed.), Delhi: National Publishing House.

[5] Hook, P. E. 1981. "Hindi Structures: Intermediate Level." Michigan Papers on South and Southeast Asia, The University of Michigan Center for South and Southeast Studies, Ann Arbor, Michigan.

[6] Butt, M. 1995. "The Projection of Arguments: Lexical and Compositional Factors", A. Alsina *et al*. (eds.), *Complex Predicates*. CSLI Publications, Stanford.

[7] Butt, M. 1993, "Conscious choice and some light verbs in Urdu." In M. K. Verma ed. (1993).

[8] Bashir, E. 1993, "Causal Chains and Compound Verbs." In M. K. Verma ed. (1993).

[9] Fedson, V. J. 1993. "Complex verb-verb predicates in Tamil." In M. K. Verma ed. (1993).

[10] Kachru, Y. 1993. "Verb Serialization in Syntax, Typology and Historical Change." In M. K. Verma ed. (1993).

[11] Pandharipande, R. 1993. "Serial verb construction in Marathi." In M. K. Verma ed. (1993).

[12] Paul, S. 2004. *An HPSG Account of Bangla Compound Verbs with LKB Implementation*, Ph.D. Dissertation. CALT, University of Hyderabad.

[13] Mohanan, T. 1995. "Wordhood and Lexicality", *NLLT* 13, 75-134.

[14] Dayal, V. 2003. "A Semantics For Pseudo Incorporation." Ms. Rutgers University

[15] Mohanan, T. 1995. "Multidimensionslity of Representation: NV Complex Predicates in Hindi", *Complex Predicates*. In A. Alsina et al.(eds).

[16] Miller,G., Beckwith,R., Fellbaum,C., Gross,D., and Miller,K. 1990. *Five Papers on WordNet.* CSL Report 43, Cognitive Science Laboratory, Princeton University, Princeton.

[17] Vossen Piek (ed.). 1998 *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers, Dordrecht.

[18] Jha Vishwanath. 1975. *Amarkosha by Amarsingha*,. Motilal Banarasidas Publications, Varanasi.

[19] Lenat D.B. and Guha R.V. 1990. *Building Large Knowledge Based System, Representation and Inference in the CYC Project*, Addison Wesley, Reading, MA. (http://www.cyc.com.)

[20] http://ontology.teknowledge.com/
[21]Chakrabarti D, Narayan D, Pandey, P and Bhattacharyya P. 2002 "Experiences in Building the Indo WordNet: A WordNet for Hindi", *Proceedings of the First Global WordNet Conference*. (http://www.cfilt.iitb.ac.in/webhwn).
[22] http://framenet.icsi.berkeley.edu/~framenet
[23] http://www.cis.upenn.edu/verbnet/
[24] Levin B., *English Verb Classes and Alternations A Preliminary Investigation,* University of Chicago Press, 1993.
[25] Dave S. and Bhattacharyya P.. 2001. "Knowledge Extraction from Hindi Texts", *Journal of Institution of Electronic and Telecommunication Engineers, vol. 18, no. 4*. pp??
[26] Rion S., Daniel J., and Ng A. Y., 2006. "Semantic taxonomy induction from heterogenous evidence". In *Proceedings of COLING/ACL 2006*. pp??
[27] Marius P., 2005. "Finding Instance Names and Alternative Glosses on the Web: WordNet Reloaded" In CICLing. pp280-292.
[28] http://www.cfilt.iitb.ac.in/wordnet/webhwn.
[29] Chakrabarti D. and Bhattacharyya, P. 2004. "Creation of English and Hindi Verb Hierarchies and their Application to Hindi WordNet Building and English-Hindi MT", Proceedings of the Second Global Wordnet Conference, Brno, Czech Republic.
[30] BNC Consortium: *British National Corpus*. 2000. The Humanities Computing Unit of Oxford University. http://www.hcu.ox.ac.uk/BNC.
[31] http://www.cfilt.iitb.ac.in/indiancorpora/web/hindi/index.html.
[32] http://wordnet.princeton.edu/.
[33] Hornby, A. S.. 2000. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford.
[34] Verma R. and Kapur B. 1998. *Lokbharti Pramanik Hindi Kosh*. Lokbharati Prakashan, ed. Varanasi,.
[35] Shree Naval Jee.(2000). *Nalanda Vishal Shabd Sagar*. Adish Book Depot, New Delhi.
[36] Mcgregor, R.S. 1997.*The Oxford Hindi English Dictionary*, Oxford university Press, New Delhi.
[37] Gupta, S.S. and Agarwal, S. 2003. *Standard Illustrated Advanced Dictionary*. Ashok Prakashan, Delhi-6.
[38] http://www.unl.ias.edu/unlsys/unl/UNL%205specifications.html
[39] Verma, M.K. (ed.) 1993. *Complex Predicates in South Asian Languages.* Manohar Publishers and Distributors, New Delhi.
[40] Alsina, A. et al. 1995. *Complex Predicates*. CSLI Publications. Stanford.