

Representing Compound Verbs in Indo WordNet

Soma Paul

International Institute of Information Technology
Hyderabad, India
soma@iiit.ac.in

Abstract

This paper proposes design for representing Indo-Aryan compound verbs in Indo WordNet. Storing these multiword expressions as a whole has been considered not a good idea because generalization will be missed and also some amount of redundancies will creep in the database. In stead we propose to set up a *lexical link* between a light verb and main verbs it combines with. The motivation of the design is triggered by the following observations: a) Light verbs are polysemous to their corresponding full verbs; however they are semantically bleached. b) CVs are lexical variants of their V1 component. By postulating a semantic relation of *Compound verbs* for V1s, we can establish the relation between CV and its V1 associate. By carefully examining the semantic nuances that light verbs add to the meaning of the main verb, the present paper also attempts to present ontology of aspect.

1 Introduction

Development of multilingual Indo WordNet is presently an ongoing endeavour in India (Ramanand et al., 2008, Sinha, 2006) which strives to construct WordNet for several Indian languages using the Hindi WordNet¹(<http://www.cfilt.iitb.ac.in/wordnet/webhwn/>) as the base. At this opportune moment it is needed that we examine various kinds of complex predicates² and determine a suitable representation for them in WordNet like online database.

Compound verb (CV) is a kind of complex predicate that frequently occurs in Indo-

Aryan languages. They are composed of two verbs, the first member – main verb (V1) – is either a participial form as in Bangla, Odiya and Assamese³ (see 1a) or a root (as in Hindi⁴, see in (1b)) and the second member – light verb (V2) – is “semantically bleached” and it bears the inflection. For example,

Bangla

- 1a. *meeTa heS-e uTh-lo*
girl-cl laugh-pf rise – 3 pt
‘The girl burst into laughter’

Hindi

- b. *laRki has paR-I*
girl laugh fall – 3 pt
‘The girl burst into laughter’

The examination of compound verbs in the context of constructing multilingual Indo WordNet is significant; the motivation being the following:

- a. The repertoire of light verbs varies from language to language⁵. For example, Bangla has

³ Bangla, Odiya and Assamese are languages spoken in eastern and north eastern zone of India. Bangla is also the official language in Bangladesh. A detailed study of compound verbs in these languages can be obtained from Dasgupta (1977), Mohanty(1992), Paul(2004).

⁴ Hindi is a widely spoken language in Northern and central India. The compound verbs of Hindi have been studied by Hook (1974), Abbi(1991), Butt(1995) to name a few.

⁵ The attestation of the compound verb is most frequent in Hindi-Urdu (Hook 1974), while it is very rare in Kashmiri (Kaul 1985). Bangla occupies the third position in the scale of frequency. Hook has conducted a contrastive typological study between the *compound-verb-rich* languages such as Hindi-Urdu and *compound-verb-poor* languages such as Marathi and concludes that the occurrence of CV sequences in *compound-verb-rich* languages have acquired a grammatical significance. In languages like Marathi the absence of the V2 has no conventional interpretation on which the hearer can rely on. In Hindi-Urdu, on the other hand, the presence of the V2, even where it is redundant, has

¹Among others, Marathi wordnet is now available online (<http://www.cfilt.iitb.ac.in/wordnet/webhwn/>)

² Indo-Aryan languages are replete with various kinds of complex predicates. For example, there exist noun+verb, adjective+verb and verb+verb constructions. Complex predicates represent one single event and are syntactically monoclausal unit (Butt (2003)).

V2 occurrence of the verb *bERano* ‘roam’ as in *bole bERano*, *kine bERano* that implies the actor is doing the action of “talking” and “buying” *randomly*, without any discretion. Hindi language shows a lexical gap for such CV. Such gaps require proper treatment during the development of multilingual Indo WordNet.

- b. Even when the same light verb occurs in two Indo-Aryan languages, their selection by main verbs might differ in those languages. For example, *pəRna* ‘fall’ occurs as a V2 both in Bangla and Hindi. However, the following CV is allowed in Hindi and not in Bangla: H⁶. *ro pəRa* – B. **keMde pORa* (B) ‘cry+fall’. In Bangla the legitimate CV is *keMde oTha* ‘burst in crying’, which is not there in Hindi even though *uThna* occurs as a V2 in that language.

WordNet presently deals with complex predicates and does include lexicalized synsets which may contain either single words or MWEs, or sometimes, both together as illustrated below from English and Hindi WordNets:

English WN {*girlfriend*, *girl*, *lady_friend*}
 Hindi WN {*uThana*, *calana*, *cheRna*, *arambh_karna*, *Suru_karna*} “start a discussion”

The present paper argues in favor of lexical status of Compound Verbs and maintains that storing compositional compound verbs as a whole in the lexicon is not a good idea because of the following reasons (Calzolari 2002):

- We lose generalizations;
- We lose the possibility to produce a proper interpretation of these constructions;
- We run into problems when operating in a multilingual environment, when something that is a MWE in a certain language has to be expressed in the target language in terms of a normal syntactic pattern.

We propose in this paper a novel design of representing compositional CVs (which can be extended for other kinds of complex predicate as

become an obligatory marker of perfectivity. Its absence has come correspondingly closer to having a conventional interpretation of imperfectivity.

⁶ H. = Hindi, B. = Bangla

well) in WordNet like lexical database. In this kind of representation, there will be no increase in the number of entries that are already present in WordNet. This will be achieved in terms of setting up a *linking* between the two components of CVs. A similar approach has been discussed in Bentvogli et al. (2004) as an alternative to *phraseset* for representing ‘recurrent free phrases’ in Italian WordNet. Hindi WordNet uses similar strategy for linking synsets across different part of speech, for example, noun – adjective. The idea of adopting the strategy of *linking* for the present task is motivated by the following observations about CVs and their component verbs:

- V2s are semantically related to their full verb counterpart, they are polysemous. Enlisting V2s under its full verb counterpart captures the polysemy factor.
- Semantic nuances contributed by a V2 to the overall meaning of CVs remains same for all cases. Those semantic features are attributed to V2 in the lexicon.
- CVs are lexical variants of their V1 component. By postulating a semantic relation of *Compound verbs* for V1s, we can establish the relation between CV and its V1 associate.

Identification of CV in a language is an issue because there are homotactic sequences that are not to be taken as compound. The next section discusses some criteria for determining a verb sequence to be CV. Section 3 examines semantic contribution of V2s. Section 4 presents the architecture of representing CV in WordNet. The issues are illustrated with Bangla data in this paper with some reference to Hindi.

2 Compound Verbs and their features

CVs are composed of more than one verb. They retain the meaning of the main verb (V1) and the V2 which is semantically bleached adds semantic nuances to the meaning of CVs. Therefore CVs are considered as lexical variant of their V1 component. On the surface, the constituent verbs enjoy a considerable amount of freedom of movement. Other syntactic elements like adverb can intervene between the constituents (The constituents are freer in Bangla compared to Hindi (see Paul 2004, Butt 1995). However, they

represent one predicate, a functional semantic unit. Adverbs and negation scope over the whole construction and cannot modify one of the components. Identifying them as representative of one predicate is the main criterion for considering them lexicalized compound and not syntactically compositional construct. The other argument in favor of the lexical status of CVs is the following: Even though CVs are lexical variants of their V1 counterparts, the argument structure⁷ of CVs is not always a copy of that of their V1 component. Nor they are licensed by their V2 constituent (for illustration see Paul 2004). On the contrary, the argument structure is licensed by the semantics of the lexicalized CV; an idiosyncratic property of the resultant construct.

In the context of designing architecture of representing compound verb repertoire in WordNet like lexicon, the discussion of “V2 semantics” becomes very significant. The semantic contribution of V2s to the meaning of the V1s creates new semantics for the CVs which can also change the syntactic behavior of the CVs. The next section will illustrate this with examples.

3 Semantics of V2s

In Bangla, we have identified 15 V2s. They are the following:

deoa ‘give’, *neoa* ‘take’, *phEla* ‘drop’, *tola* ‘lift’, *rakha* ‘keep’, *oTha* ‘rise’, *pORa* ‘fall’, *bERano* ‘roam’, *bOSa* ‘sit’, *aSa* ‘come’, *jaoa* ‘go’, *cOla* ‘move’, *ana* ‘bring’, *mOra* ‘die’, *paThano* ‘send’.

All V2s have full verb counterparts in the language with which they share the core meaning⁸. As light verbs, they undergo semantic loss. Hook (1974) has taken an extreme position and states that V2s becomes lexically empty⁹. Others have

⁷ Argument structure is an ordered list associated to lexical representation of a verbal predicate that contains arguments licensed by that predicate.

⁸ Paul [2004] has studied the concept of core meaning in detail and how core meaning is shared by the V2s and their corresponding full verb.

⁹ Hook describes the phenomenon as *grammaticalization* (Hook 1974:94-97). Sarkar (1975) elucidates Porizka’s perception of grammaticalization – a stripping off of the main dictionary meaning from the vector verb in order to reduce them to the role of ‘aspective’.

conferred a semi-lexical status to V2s¹⁰. Butt (1995) assumes that the light verb use of a verbal item and its use as a full verb should be identified as a case of lexical polysemy and not as grammaticalization. I will adopt Butt’s position and examine the semantic contribution of V2s in this section.

We maintain that semantics of V2 determines how the event structure of the main verb is profiled¹¹ or focused as it unifies with a V2. The profiled segment constitutes the meaning of the CV. The profiling is accomplished at two levels:

- a. By highlighting the *manner of involvement* of the participant(s) engaged in the base-event; and
- b. By imposing *temporal* and *aspectual* focus on the event denoted by the resultant CV predicate

I will describe the two levels in the following subsections.

3.1 Manner of Involvement of the participants

Some V2s profile manner of involvement of participants engaged in action denoted by V1 component and the profiled information constitutes the semantics of CV. We will examine the issue with respect to the two V2s: *deoa* ‘give’ and *neoa* ‘take’. The V2 *deoa* ‘give’ specifies that the effect of the action denoted by the main verb is directed towards a participant other than the actor; while the V2 *neoa* ‘take’ entails that the result is directed to the actor himself/herself. This amounts to understanding how participants are affected by the result of the action. I call this *semantics of affectedness*. The semantics of many verbs are inherently marked for *semantics of affectedness*. For example, the meaning of the following verbs *gOchano* ‘foist something on somebody’, *oSkano*

¹⁰There is a great deal of discussion available in the literature regarding the semantics of V2. Following are some references related to the works on Indo-Aryan Compound verb structure: Hook (1974), Sarkar (1975), Dasgupta(1989), Abbi(1991,1992), Bashir(1992), Mohanty (1992), Butt(1995), Paul(2004).

¹¹ The theory of *profiling* gives an account of constituting the meaning of compound verb is proposed in detail in Paul (2005).

‘instigate’, *dabano* ‘suppress’, *goMtano* ‘thrust’, *bigRono* ‘spoil’, *mara* ‘kill’ entails that the result of the action is directed towards an affected entity who is not the doer. These verbs can therefore select the V2 *deoa* ‘give’ and they are incompatible with V2 *neoa* ‘take’. On the other hand, verbs such as *bhaba* ‘think’, *khaoa* ‘eat’, *Sekha* ‘learn’, *paoa* ‘get’, *bojha* ‘comprehend’, *dEkha* ‘see’, *Sona* ‘hear’ entail that the doer himself is the affected entity – the recipient of the result of the action. They select V2 *neoa* ‘take’. There exist verbs which are not inherently marked for *semantics of affectedness*. They are accomplishment verbs such as *banano* ‘build’, *kena* ‘buy’, *khoMja* ‘search’ *raMdha* ‘cook’, *bhaja* ‘fry’, *Tañano* ‘hang up’, *kaTa* ‘cut (vegetables)’, *aMka* ‘draw’, *ana* ‘bring’, *kOra* ‘do’ and so on that denotes a situation in which an actor performs some action and the action has a natural outcome. There exists a culmination which borne a result of the action. The semantics of these verbs, however, does not specifically indicate to whom the result of action is directed. For example, in case of the verb *banano* ‘build’ the builder can build a house for himself (as demonstrated in (3a)) or he can build a house for

the benefit of a receiver or beneficiary as shown in (3b):

3. *binu nije-r jonne/ritu-r jonne*
 Binu self-gen for Ritu-gen for
baRi-Ta bana-len
 house-cl build-3 hon pt
 a. ‘Binu has built the house for himself’
 b. ‘Binu built the house for Ritu’

For such verbs the V2s *deoa* ‘give’ and *neoa* ‘take’ remove this vagueness by categorically focusing on the *manner* in which participants are involved in a situation. The CV *banie neoa* “build-cp take” profiles the self-directedness (or self-beneficiary) reading inherent in the semantics of the verb *banano* ‘build’. The CV *banie deoa* “build-cp give”, on the other hand, specifies that the effect of the action directed towards an entity other than the doer. Ritu is the beneficiary in the following sentence:

4. *binu ritu-ke Ek-Ta baRi bani-e*
 Binu Ritu-obj one-cl house build-cp
*di-lo /*ni-lo*
 give-3pt/ take-3pt
 ‘Binu built a house for Ritu’

The following table presents the above discussion:

V1	CV with V2 <i>deoa</i> ‘give’	Semantic overtone	CV with V2 <i>neoa</i> ‘take’	Semantic Overtone
<i>Banano</i> ‘build’	<i>banie deoa</i> ‘build a thing (for someone)’	Effect of the action directed towards a participant other than the actor	<i>banie neoa</i> ‘build a thing (and the benefit goes to the doer)’	Effect of the action directed towards the actor himself (self-beneficiary)
<i>Bhaba</i>	* <i>bhebe deoa</i>		<i>bhebe neoa</i> ‘think within oneself’	Effect of the action directed towards the actor himself (self-beneficiary)
<i>gOchano</i>	<i>gochie deoa</i>	Effect of the action directed towards a participant other than the actor	* <i>gochie neoa</i>	

Table 1: Manner of Involvement of participants in action denoted by CVs

The above data illustrates how the semantics of V2 profiles the participant role involved in the base

structure of the main verb component. The next sub-section substantiates the claim that V2s adds

temporal and aspectual focus to the CVs which are lexicalized.

3.2 Telicity and Duration as Inherent Property of CV's Semantics

Scholars in recent years (Vendler 1967, Smith 1991 among others) no longer perceive aspectual notions such as *duration* and *telicity* as an entirely grammaticized concept. Vendler's (1967) classification of verbs into accomplishment, achievement, activity and stative effectively includes *duration* and *endpoints of events* as an integrated part of the semantics of verbs. Carlota Smith has used the concept of

aspect in a broader sense and identifies various lexical spans for verbs event structure. For example, "The verb constellation 'arrive in Boston' spans a moment near the end of a chain of events while 'go to Boston' covers a much larger part of chain" (Smith 1991, p 34). We maintain that verbs represent an event structure. The straight line in the following figure indicates the event line. The broken line on left and right sides of the straight line indicates the period prior to the starting of the event and the resultant state respectively. We have attempted to present how V2s profiles lexical span of V1 in the following diagram. The profiled segment determines the aspect of the resultant CV:

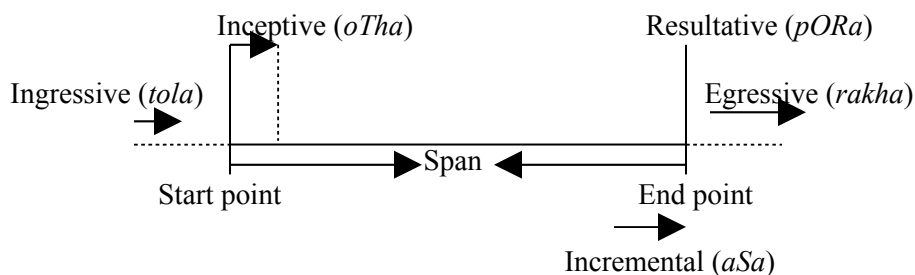


Figure 1: Lexical span profiled by V2s

Inceptive: focus on the entry into an event and thus spans the early portion of the causal chain (Measured by the broken vertical line)

Ingressive: time elapsed before the beginning of the situation¹²

Egressive: focus on the exit from an event and spans the later portion

Resultative: focus on the end of the causal chain

Incremental: focus on the developmental span as the event progress towards its final outcome

Imminent: Approaching towards a goal (not necessarily a developmental span)

Along with the aspects presented in Figure 1, we have more well-known aspectual viewpoints: imperfective and perfective. These aspects have two basic lexical semantic features: *telicity* and *duration*. This section attempts to make evident that these features are inherent of CVs and they are contributed by V2s to the overall

meaning of CVs. For example, the main verb *kaMda* 'cry' is an activity which entails duration. When this verb combines with the V2 *phEla* 'drop', the resultant CV becomes non-durative in nature. However, the durative nature prevails when the verb selects V2s such as *cOla* 'move', *bERano* 'roam'. The adverb *EknagaRe* 'at a stretch' conveys duration. As shown in the following examples, this adverb is compatible with *kaMda* 'cry', *keMde cOla* 'continue to cry' and not with *keMde phEla* 'cry unpremeditatedly':

5a. *ritu EknagaRe kaMd-che / keMd-e col-eche*
Ritu at a stretch cry-3 pr ct cry-cp move-3 pr ct
'The girl is laughing continuously'

5b. * *mee-Ta EknagaRe keMd-e phelche*

kaMda 'cry' is an atelic verb because it does not include an endpoint. However, when this verb occurs with *oTha* 'rise', the resultant CV becomes telic in nature and become compatible with completive adverbials as illustrated below:

¹² Smith illustrates that the completive adverbial phrase 'in an hour' in the sentence 'He left the house in an hour' refers to the time interval at the end of which the event of 'leaving the house' takes place.

6. *ritu muhurt-er moddhe keMd-e uTh-lo*
 Ritu second-gen in cry rise-e pt
 ‘Ritu burst into crying within a moment’

Maraffa (2003) has proposed a solution to represent lexical telicity in WordNets-like

computational lexica for Portuguese telic complex predicate. Besides aspectual features, the other lexical property that the CVs inherently reflect is modal information. The following table presents the overall semantic contribution that V2s add to build the semantics of CVs:

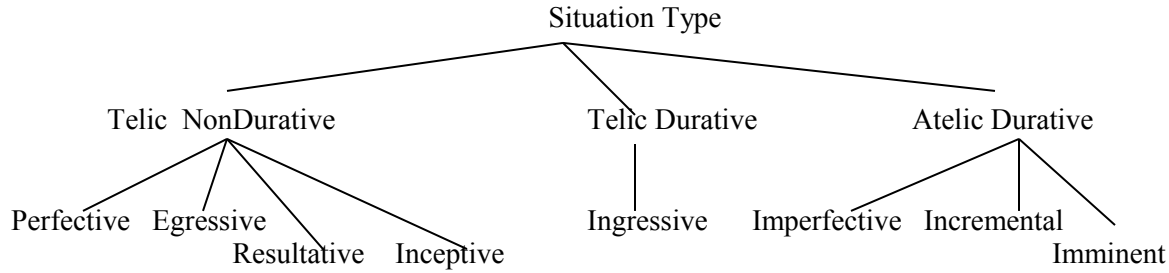
	V2	Aspect	Mood	Telicity	Duration	Participant
1	<i>deoa</i> ‘give’	Perfective		+	-	Non-self
2	<i>neoa</i> ‘take’	Perfective		+	-	Self
3	<i>phEla</i> ‘drop’	Perfective	Unpremeditated	+	-	Volitional
4	<i>tola</i> ‘lift’	Ingressive		+	+	
5	<i>rakha</i> ‘keep’	Egressive		+	-	
6	<i>oTha</i> ‘rise’	Inceptive	Suddenness	+	-	Upward
7	<i>pORa</i> ‘fall’	Resultative	Immediateness	+	-	Downward
8	<i>bERano</i> ‘roam’	Imperfective		-	+	
9	<i>bOSa</i> ‘sit’	Resultative	Unwarranted	+	-	
10	<i>aSa</i> ‘come’	Incremental		-	+	
11	<i>jaoa</i> ‘go’	Imperfective		-	+	
12	<i>cOla</i> ‘move’	Imperfective		-	+	
13	<i>ana</i> ‘bring’	Imminent		-	+	
14	<i>mOra</i> ‘die’	Imperfective	Futility	-	+	
15	<i>paThano</i> ‘send’	Perfective		+	-	

Table 2: Semantics of V2s

The following section attempts to present ontology of aspect. All situation types can be represented in this ontology and this will be applicable for all verbs, simple or compound.

3.3 Ontology of aspect

The most abstract ontological category in the following ontology is called *Situation Type*. For the V2s, we postulate the following situation types which subsume various aspects.



4 Representation of CVs in WordNet

Unlike English WordNet (1998), where phrasal verbs are listed under the main verb, we propose to organize V2s as a synset to their full verb counterpart. Compound verb is specified as a link between V1 and V2. The strategy is very similar to the one that Hindi WordNet is already using for cross part of speech linkage. Hindi WordNet has adopted a device of cross parts of speech linkage by which relations between the synsets of different part of speech is established. For example, the nominal and verbal concepts are linked by *ability link*, *capability link* and *function link*. The following example illustrates *function link* that describes the ‘function’ or *karma* of the noun referred to by *adhyaapak*.

अध्यापक, शिक्षक, आचार्य, गुरु, मास्टर (adhyaapak, shikshak, aacaarya, guru, master; *teacher*)

==> पढ़ाना, शिक्षा देना (paRhaanaa, shikshaa denaa; *teach*)

For compound verbs, linking is established between the main verb and V2 synsets. The link will indicate the meaning of the resultant CV. For every main verb, there will be a link “Compound Verb” which will specify all V2s that the main verb can combine with. Let us illustrate the proposal with an example. The verb *ghumono* ‘sleep’ can occur with the following V2s (see column 2):

V1	V2	CV	Meaning
<i>ghumono</i>	<i>pORA</i>	<i>ghumie pORA</i>	Fall asleep
<i>ghumono</i>	<i>neoa</i>	<i>ghumie neoa</i>	Take a nap

The information related to the entry *ghumono* will be the following:

Verb

(R)*ghumono* – *nidrar Oboshthay thaka* ‘to be in a state of sleep’

- A. Ontology Node
- B. Hypernymy
- C. Compound Verb
- ...

Once the link for ‘Compound Verb’ is expanded the V2s and their semantics will be displayed as illustrated below:

Verb

(R)*ghumono* – *nidrar Oboshthay thaka* ‘to be in a state of sleep’

- A. Ontology Node
- B. Hypernymy
- C. Compound Verb

* (R) *neoa* – Result of the action is directed to the actor

- A. Ontological Node

- * Perfective
 - * Telic Non-Durative
 - * Situation Type

* (R) *pORa* – Action done with an immediateness effect

- A. Ontological Node

- * Resultative
 - * Telic Non-Durative
 - * Situation Type

The advantage of this kind of representation is linguistically very significant. First, we can assert

through the design that V2s and their full verb counterparts are polysemous. They are semantically related. Second, the relation “Compound Verb” on main verb signifies that compound verbs are lexical variant of their main verb counterpart. We will also be able to present all CVs at one place which are lexical variants of a main verb. Third, the design helps us to state that a V2 adds same semantic nuance whenever it unifies a main verb.

5 Conclusion

Compound verbs are viewed as lexicalized items in this work. This paper presents a design of representing compound verbs in Indo WordNet. The proposal is to set up *lexical link* between main verb and the V2s that it selects. The V2s with their gloss is listed as polysemy to their corresponding full verb entry. Ontology of aspect is built to represent the semantic import of the V2s. The present study is done with respect to Bangla data. The task that remains is to handle compound verbs in multilingual scenario with a view of organizing them in Indo WordNet like database.

Reference

- Anvita Abbi. 1991. “Semantics of Explicator Compound Verbs.” In *South Asian Languages, Language Sciences*, volume 13:2, 161-180.
- Carlota Smith. 1991. *The Parameter of Aspect*. Kluwer Academic Publishers, The Netherlands.
- C. Fellbaum et al. 1998. *Wordnet: An Electronic Lexical Database*. MA: The MIT Press.
- D. Chakrabarti, D. Narayan, P. Pandey and P. Bhattacharyya. 2002. Experiences in Building the Indo WordNet – A WordNet for Hindi. In *Proceedings of the First International Conference on Global WordNet (GWC02), Mysore, India*.
- G. Mohanty. 1992. *The Compound Verbs in Oriya*. Ph. D. dissertation, Deccan College Post-Graduate and Research Institute, Pune.
- J. Ramanand, A Ukey, B.K.Singh, P. Bhattacharyya. 2008. *Mapping and Structural Analysis of Multi-Lingual WordNets*. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering
- L. Bentivogli and E Pianta. 2004. Extending WordNet with syntagmatic Information. In *Proceedings of Second Global WordNet Conference*.
- M. Sinha, M. Reddy, P. Bhattacharyya. 2006 *An Approach Towards Construction and Application of Multilingual Indo-WordNet*. Proceedings of the 3rd Global WordNet Conference (GWC 05), Jeju Island, Korea.
- Miriam Butt. 1995. *The Structure of Complex Predicates in Urdu*. Doctoral Dissertation, Stanford University.
- N. Calzolari, Fillmore C., Grishman R., Ide N., Lenci A., MacLeod C., Zampolli A. 2002 Towards Best Practice for Multiword Expressions in Computational Lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, (1934-1940).
- Palmira Marrafa. 2005. The Representation of Complex Telic Predicates in WordNets. In J. Cardenosa et al. *Universal Network Language: Advance in Theory and Application*.
- Peter Hook. 1974. *The Compound Verbs in Hindi*. The Michigan Series in South and South-east Asian Language and Linguistics. The University of Michigan.
- Probal Dasgupta. 1977. The internal grammar of compound verbs in Bangla. *Indian Linguistics*, 38(3):68-85.
- S. Paul. 2004. An HPSG Account of Bangla Compound Verbs with LKB Implementation. Ph.D dissertation, University of Hyderabad, Hyderabad.
- S. Paul. 2005. The semantics of Bangla Compound Verbs. *Yearbook of South Asian Languages and Linguistics*. 101-112.
- Vijay K. Kaul. 1985. *The Compound Verb in Kashmiri*. Unpublished Ph.D. dissertation. Kurukshetra University.