# Hindi Compound Verbs and their Automatic extraction

**Debasri Chakrabarti**
Humanities and Social Sciences Department
IIT Bombay
debasri@iitb.ac.in

**Vaijayanthi M. Sarma**
Humanities and Social Sciences Department
IIT Bombay
vsarma@iitb.ac.in

**Pushpak Bhattacharyya**
Computer Science and Engineering Department
IIT Bombay
pb@cse.iitb.ac.in

## Abstract

We analyse Hindi complex predicates and propose linguistic tests for their detection. This analysis enables us to identify a category of V+V complex predicates called *lexical compound verbs* (*LCpdVs*) which need to be stored in the dictionary. Based on the linguistic analysis, a simple automatic method has been devised for extracting *LCpdVs* from corpora. We achieve an accuracy of around 98% in this task. The *LCpdVs* thus extracted may be used to automatically augment lexical resources like wordnets, an otherwise time consuming and labour-intensive process

## 1 Introduction

Complex predicates (CPs) abound in South Asian languages [Butt, 1995; Hook, 1974] primarily as either, noun+verb combinations (*conjunct verbs*) or verb+verb (V+V) combinations (*compound verbs*). This paper discusses the latter.

Of the many V+V sequences in Hindi, only a subset constitutes true CPs. Thus, we first need diagnostic tests to differentiate between CP and non-CP V+V sequences. Of the CPs thus isolated, we need to distinguish between those CPs that are formed in the syntax (derivationally) and those that are formed in the lexicon (*LCpdVs)* in order to include only the latter in lexical knowledge bases. Further, automatic extraction of *LCpdVs* from electronic corpora and their inclusion in lexical knowledge bases is a desirable goal for languages like Hindi, which liberally use CPs.

This paper discusses Hindi Verb+Verb (V+V) CPs and their automatic extraction from a corpus.

### 1.1 Related work

Alsina (1996) discusses the general theory of complex predicates. Early work on conjunct and compound verbs in Hindi appears in Burton-Page (1957) and Arora (1979). Our work on diagnostic tests for CPs, as reported here, has been inspired by Butt (1993, 1995 for Urdu) and Paul (2004, for Bengali). The analysis of lexical derivation of *LCpdVs* derives from the work on compound verbs by Abbi (1991, 1992) and Gopalkrishnan and Abbi (1992).

This work is motivated primarily by the need to automatically augment lexical networks such as the Princeton Wordnet (Miller *et. al.,* 1990) and the Hindi Wordnet (Narayan *et. al.*, 2002). Pasca (2005) and Snow et. al. (2006) report work on such augmentations by processing web documents.

To the best of our knowledge ours is the first attempt at automatic extraction of *LCpdVs* from Hindi corpora.

### 1.2 Organization of the paper

Section 2 discusses CPs in Hindi and the ways to distinguish them from other, similar looking, constructions. Section 3 discusses the automatic extraction of CPs from corpora. Section 4 concludes the paper.

## 2 V+V Complex Predicates in Hindi

We have identified five different types of V+V sequences in Hindi. These are:

**1. V1 stem+V2:** *maar Daalnaa* (kill-put) 'kill'.

**2. V1 inf-e+lagnaa:** *rone lagnaa* (cry-feel) 'start crying'.
**3. V1 inf+paRnaa:** *bolnaa paRaa* (say-lie) 'say'.
**4. V1 inf-e+V2:** *likhne ko/ke lie kahaa* 'asked to write'.
5**.** V1–kar+V2**: *lekar gayaa* 'took and went'.**

### 2.1 Identification of CPs]

Following Butt (1993) and Paul (2004), we use the following diagnostic tests to identify CPs in Hindi:

1. Scope of adverbs
2. Scope of negation
3. Nominalization
4. Passivization
5. Causativization
6. Movement

(see Appendix A for an example of these tests)

The tests above have been exhaustively applied to varied data. The results of these tests show that some V+V sequences function as single semantic units and others do not. They also show that the **V1stem+V2**, **V1inf-e+lagnaa** and **V1inf+paRnaa** sequences show similar properties and the **V1 inf-e+V2** stem and the **V1–kar+V2** behave similarly. We call these Group 1 and Group 2 respectively.

Group 1 sequences are true CPs in Hindi. The V+V sequences are simple predicates (monoclausal) with one subject. Group 2 constructions are not CPs. They show clausal embedding and each verb behaves as if it were an independent syntactic entity. In the next section we summarize the semantic properties of CPs (Group 1).

### 2.2 Semantic Properties of V2 in Group 1

After identifying the CPs from among different V+V sequences, the next step was to determine how they are formed. To accomplish this we examined the semantic properties of the second verbs (V2) in Group 1:

**(1) V1inf+paRnaa:**
Examples include *karnaa paRaa* 'do-lie (had to do)', *bolnaa paRaa* 'say-lie (had to say)' etc. The second verb is always *paRnaa* 'to lie (lay)'. It appears in its stem form and bears all the inflections. As V2, *paRnaa* has the meaning of *compulsion/force. paRnaa* 'lie' as a V2 can be combined with any V1 irrespective of the latter's semantic properties. Since there are no syntactic or semantic restrictions on the selection of V1, this construction should be treated in the syntax as a combination of a V1 and a modal auxiliary.

**(2) V1 inf-e+lagnaa:**
Examples include *karne lagaa* 'do-feel (start to do)', *bolne lagaa* 'say-feel (start to say)' etc. The V2 in this sequence is always *lagnaa* 'feel' in the bare form and carries all the inflections. The core meaning of *lagnaa* 'feel' is lost when it is combined with a V1. As a V2 it always has the meaning of *beginning, happening of an event*. *lagnaa* 'feel' as a V2 can be combined with any V1 irrespective of the latter's semantic properties. Thus, this is also an instance of a modal auxiliary and should be derived in the syntax.

**(3) V1stem+V2**
In the formation of V1 stem+V2, the V2 may be any one of ten verbs, as shown in Figure 1.

1. **Daalnaa '*put*'**
2. **lenaa '*take*'**
3. **denaa '*give*'**
4. **uThnaa '*wake*'**
5. **jaanaa '*go*'**
6. **paRnaa '*lie*'**
7. **baiThnaa '*sit*'**
8. **maarnaa '*kill*'**
9. **dhamaknaa '*throb*'**
10. **girnaa '*fall*'**

**Figure 1: The 10 vector verbs**

All these V2s also occur as main verbs. As V2, the core meaning of these verbs is lost (bleached), but they acquire some new semantic properties which are otherwise not seen (Abbi, 1991, 1992; Gopalkrishnan and Abbi, 1992). The semantic properties of V2s include *finality, definiteness, negative value, manner of the action, attitude of the speaker etc.*

The combination of V1 and V2 is subject to the semantic compatibility between the two verbs. The argument structure of the CP is determined by V1 as is the case-marking on the internal arguments, but the case-marking on the external argument (subject) is determined by both verbs.

From this analysis we conclude that V+V CPs are formed both lexically and syntactically in Hindi. Detailed investigation shows us that the V2 in the **V1inf-e+lagnaa** and the **V1inf+paRnaa constructions** is a type of modal auxiliary and its semantic features are predictable and unvarying. We propose to deal with these verbs in the syntax and call these verbs *syntactic compound verbs* (*SCpdVs*). The V2 choice in the V1stem+V2 is not predictable and the CPs func-

tion as a single complex of syntactic and semantic features. We call these verbs *lexical compound verbs* (***LCpdVs***) and we propose to include them in the lexical knowledge base. In the next section we provide a heuristic for automatic extraction of *LCpdVs* for storage in the lexicon.

### 2.3 The Extraction Process

By scanning the corpus, V1stem+V2 sequences were found given the heuristic *H** specified in Figure 2.

**(Heuristic H*)**
**If a verb V1 is in the stem form and is followed by a verb V2 from a pre-stored list of verbs that can form the second component of the CP (section 2.2, Figure 3), i.e., the 'vector', then this verb along with the V2 is taken to be an instance of an LCpdV.**

**Figure 2: Main heuristic for identifying *LCpdVs***

Ten native speakers of Hindi were consulted. They were asked to construct sentences with the extracted sequences. If they were able to do so, that sequence was registered as a true *LCpdV*.

The precision of the heuristic is calculated as the ratio of the *actual LCpdVs* arrived at through manual validation to the total number of *anticipated LCpdVs* identified by the heuristic.

The results of these calculations are shown in Table 1, with a precision rate of 70% for the BBC corpus and 79% for the CIIL one.

| **Corpus** | **Total detections** | **POS ambiguities** | **Passive forms** | ***LCpdV*s (manually detected)** | **Precision** |
|---|---|---|---|---|---|
| BBC | 40 | 8 | 4 | 28 | 0.7 (28/40) |
| CIIL | 174 | 32 | 7 | 135 | 0.79 (135/174) |

**Table 1: Precision of *LCpdV* extraction**

The loss in precision was caused by (i) part of speech ambiguity, (ii) passivisation and (iii) idiomatic usages. For lack of space, we discuss only (i) here. Consider the following example:

**(*bhaag* 'part (N) / run away (Vb))':**

In Hindi, *bhaag* can be both a noun meaning 'part', as in

*vah is khel mẽ* ***bhaag liaa***
*'He took part in this game.'*

and a verb *bhaagnaa* 'flee', as in
*jel se do kaedii* ***bhaag gaye***
*'Two prisoners ran away from the jail.'*

The automatic system for *LCpdV* identification mistakenly flags the sequence *bhaag lenaa* (take part) as one, simply because *bhaag* looks like the stem form of the verb *bhaagnaa* and V2 is from the pre-constructed list of vectors (Figure 1). A sentence like (0) is ungrammatical in Hindi.

(1) **jel-se do kaedii bhaag lie*
*From jail-loc two prisoner escape take-perf-pl*
*'Two prisoners escaped from the jail'.*

The verb *bhaagnaa* 'escape' as a V1 never selects *lenaa* 'take' as a V2 to form an *LCpdV*. The noun/verb POS ambiguity results in incorrect identification of *LCpdV*s.
When measures were taken to remedy these errors, we reached an accuracy of close to 98% (see table 2).

| | BBC | CIIL |
|---|---|---|
| Confirmed *LCpdVs* (A) | 423 | 953 |
| Not *LCpdVs* (B) | 13 | 12 |
| Different POS (C) | 65 | 179 |
| Possible *LCpdVs* but contexts insufficient (D) | 44 | 36 |
| Minimum Precision (A/(A+B+D)) | 0.88 (423/480) | 0.95 (953/1001) |
| Maximum Precision ((A+B)/(A+B+D)) | 0.97 (467/480) | 0.99 (989/1001) |
| Total V1stem+V2 constructions in the corpus | 10,145 | 36,115 |

**Table 2: Final results of LCpdV extraction**

A partial list of *LCpdVs* extracted from a test run on the CIIL corpus is presented in Table 3.

| | | | | |
|---|---|---|---|---|
| baandh denaa 'tie' | Kar lenaa 'do' | Bhar denaa 'fill' | le jaanaa 'take' | Banaa denaa 'make' |
| jaan lenaa 'know' | kaaT denaa 'cut' | Kar denaa 'do' | Badal jaanaa 'change' | Bhuul jaanaa 'forget' |

| | | | | |
|---|---|---|---|---|
| jalaa denaa 'burn' | Gir jaanaa 'fall' | Samajh lenaa 'under-stand' | Samjhaa denaa 'make under-stand' | Khod lenaa 'dig' |
| lauTaa denaa 'return' | Rah jaanaa 'stay' | Le lenaa 'take' | De denaa 'give' | ghusaa denaa 'enter' |

**Table 3: Examples of LCpdV extraction**

## 3 Conclusions and Future Work

In this paper, we have presented a study of Hindi compound verbs, proposed diagnostic tests for their detection and given automatic methods for their extraction from a corpus. Native speakers verify that the accuracy of our method is close to 98% on representative corpora.

Future work will consist in inserting the extracted *LCpdV*s into lexical resources such as the Hindi wordnet[2] at the right places with the right links.

## References


Abbi, Anvita. 1991. *Semantics of explicator compound verbs*. In South Asian Languages, Language Sciences, 13:2, 161-180

.Abbi, Anvita. 1992. *The explicator compound verb: some definitional issues and criteria for identification*. Indian Linguistics, 53, 27-46.

Alsina, Alex. 1996. *Complex Predicates:Structure and Theory.* CSLI Publications,Stanford, CA.

Arora, H. 1979. *Aspects of Compound Verbs in Hindi.* M.Litt. dissertation, Delhi University.

Burton-Page, J. 1957. *Compound and conjunct verbs in Hindi*. BSOAS 19 469-78.

Butt, M. 1993. *Conscious choice and some light verbs in Urdu*. In M. K. Verma ed. (1993) Complex Predicates in South Asian Languages. Manohar Publishers and Distributors, New Delhi.

Butt, M. 1995. *The Structure of Complex Predicates in Urdu.* Doctoral Dissertation, Stanford University.

Cruys Time De and B. V. Moiron. 2007. *Semantics-based multiword expression extraction*. ACL-2007 Workshop on Multiword Expressions.

Gopalkrishnan, D. and Abbi, A. 1992. *The explicator compound verb: some definitional issues and criteria for identification*. Indian Linguistics, 53, 27-46.

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

http://www.cogsci.princeton.edu/~wn

Narayan, D., D. Chakrabarty, P. Pande, and P. Bhattacharyya. 2002. *An experience in building the Indo WordNet - a WordNet for Hindi*, International Conference on Global WordNet (GWC 02), Mysore, India, January.

Pasca, Marius, 2005**.** *finding instance names and alternative glosses on the web: WordNet reloaded.* Proceedings of CICLing, Mexico City.

Snow, Rion, Dan Jurafsky, and Andrew Y. Ng. 2006. *Semantic taxonomy induction from heterogenous evidence*. Proceedings of COLING/ACL, Sydney.


### Appendix A. Example of a diagnostic Test for *LCpdVs:* scope of adverbs

| Verb Type | Example | Comment | CP? |
|---|---|---|---|
| V1 stem+ V2 | us-ne jaldii jaldii khaa li-aa'(S)he ate quickly.' | Scope over the whole sequence | Yes |
| V1inf-e+ lag-naa | vah jaldii se khaan-e lag-aa 'He started eating immediately.' | Scope over the whole sequence | Yes |
| V1 inf+ paRnaa | mujhe yah kaam jaldii karnaa paR-aa 'I had to do the work quickly.' | Scope over the whole sequence | Yes |
| V1inf-e+V2 | us-ne mujhe khat jaldii se likhn-e kah-aa 'He asked me to write the letter quickly.' | Either over V1 or V2 depends upon the syntactic position of the adverb | No |
| V1–kar+ V2 | vah jaldii se nahaa-kar aa-yeg-aa 'He will take bath quickly and come.' | Either over V1 or V2 depends upon the syntactic position of the adverb | No |

[2] Developed by the wordnet team at IIT Bombay, www.cfilt.iitb.ac.in/webhwn