Interdisciplinary Journal of Linguistics
Volume [15] 2022, pp. 75-84

FINDING CVs: CHALLENGES OF A CORPUS-BASED APPROACH

Mona Parakh[*]

Abstract

The aim of this study is to examine compound verbs (CVs) in Gujarati using a corpus-based methodology by way of identifying combinations of verbal sequences in a POS tagged Corpus. Empirical observations about CVs, their structures, and their occurrence can be made using the corpus, thereby enabling the formulation of morpho-syntactic and syntacto-semantic rules that are verifiable. A CV is generally made via the combination of verbs with other verbs.

The present discussion is centered around Compound Verbs (CVs) in Gujarati, focusing on their structure and their patterns of occurrence, as well as the challenges faced in the course of identifying/extracting CVs from a monolingual POS tagged corpus of Gujarati. This paper is a discussion of part of a larger research problem aimed at studying and analysing complex predicates in Gujarati, using electronic corpora. The corpus used for this study is annotated using a broad annotation scheme which labels the morpho-syntactic features in a fair amount of detail.

Keywords: Complex Predicates, Corpus-Based, Compound Verbs, Gujarati

1. Complex Predicates

A multi-word compound that functions as a single verb is a complex predicate (Das P.K. 2006). Complex predicates (CP) exist in great numbers in South Asian languages and are formed by a combination of nouns, adjectives and verbs with other verbs. The verb in the CP is referred to as light verb and the element that the light verb combines to form a CP is referred to as a host.

In case of CPs, Noun/Adjective+Verb combinations are called conjunct verbs and Verb+Verb combinations are called compound verbs. Examples of conjunct verbs in Gujarati are provided in (1) and (2) while an example of a Compound verb is provided in (3) below.

(1) Noun+Verb

maja: a:vavi

[*] Maharaja, Savagirao University of Baroda, Vadodara, India.

fun come-inf-FS

'to have fun/ to enjoy'

(2) Adjective+Verb

si:dhu karvuM

straight do-inf

'to straighten'

(3) Verb + Verb

boli: padvuM

speak-nf fall-inf

'to speak out' (suddenly)

In compound verbs the main verb (v1), also called the 'polar verb' contains the central meaning of the complex verb-form. The second verb (V2) of the sequence is a 'vector verb, which is semantically de-lexicalised/bleached or grammaticalised, so that it does not preserve its lexical meaning and functions as an auxiliary verb. It explicates the meaning of the polar or main verb and so the compound verb is also called an 'Explicator Compound Verb' (Das P.K. 2006).

2. Structure of Compound Verbs in Gujarati

For almost every CV there is a corresponding simple verb (SV) without the auxiliary, wherein the main verb carries the inflectional suffix. A simple verb such as *samajyo* understand-PST.M.SG 'understood' has a corresponding CV *samji: gayo* understand-PSTPFV.M.SG 'fully understood'. The Gujarati compound verb (CV) is composed of the non-finite form of a main or primary verb followed by the inflected form of an auxiliary or vector verb as seen in example (3). These latter verbs are homophonous with members of Gujarati's inventory of primary verbs. When used as primary verbs they express some change in location, posture or actions. As vectors they express subtle nuances of aspect and speaker attitude (Hook 1995). The semantic properties of V2s include finality, definiteness, negative value, manner of the action, attitude of the speaker etc. (Bhattacharya et al. 2008). Non-finite main verbs in Gujarati take the form such as V + /i:/. Take for example, the non-finite forms of the main verbs kha: 'eat', jo 'look', bol 'spea'' which are khai:, joi:, boli:.

3. Patterns of Occurrence of CVs in Gujarati

Hook (1974) identifies a number of environments that guarantee the occurrence of CVs in Hindi and Urdu. Two of these environments, he finds have a favoring effect on the appearance of CVs in Gujarati too. These environments include clauses which explicitly mark one event as anterior to another and complement clauses of predicates expressing fear or anxiety. Examples (4) and (5) extracted from the Gujarati Corpus substantiate the above claim.

(4) em kehvaatu hatu ke Dhakani: malmalno tako vĩ:timathi: pasa:r **thayi jato**.

so-QUOT say-CAUS.PSTPTCP that-COMP Dhaka-GEN.F muslin-GEN.M bundle-MS ring-LOC.ABL pass happen-nf go-MS-PST.

'It was said that a bundle of muslin from Dhaka would tend to pass through a ring.'

(5) ungh nahi aave eva bhaythi unghni goLini akhi shi:shi: oshika niche mukto ane *darine* ungh **avi: jati:**

sleep-FS NEG come-FUT such-DET fear-CAUS sleep-GEN-FS medicine-GEN-FS whole-Q bottle-FS pillow-OBL below keep-HAB-MS and fear-NF sleep-FS **come-NF go-FS-PST.**

'Scared that sleep would evade me, I would keep the whole bottle of sleeping pills under the pillow and out of fear I would fall asleep.'

We find that in example (4), the complementiser *ke* 'that' explicitly marks the subordinate clause as having an event anterior to the main clause, giving rise to the CV *thayi jato* happen-PSTPFV.M.SG 'tend to'. In example (5), the complement clause of the predicate *darine* 'due to fear' has the CV *av:i jati:* come-PSTPFV.F.SG 'would come over'.

4. About CVs in Other Languages

Paul (2003) discussing Bangla CV sequences using a constraint-based mechanism within HPSG framework claims that it is the V1 (and not the V2 as is generally assumed in case of Complex Predicate composition) that selects a V2. Both V1 and V2 in a CV structure have semantic content, so the unification of V1 and V2 takes place at the semantic level. The combinatorial well-formedness of a CV structure depends on the semantic compatibility between V1 and V2. According to Paul (2004),

vector verbs in CVs share with their corresponding full verb a core meaning and the relation between them is identified as one of Polysemy. Unlike the auxiliaries which are completely stripped of the core sense and function as pure grammatical categories, the vector verbs add semantic nuances to the predicate.

Hook (1974) says that vector verbs are fully emptied of their lexical content. They are grammaticalised; as such the occurrence of any given vector, as opposed to its absence, does not so much alter the meaning as the presence of their homonymous counterparts among the main verb would in a sentence.

Dasgupta (1977) says that a compound verb contains only one main verb or polar verb. The vector is a minor verb in that it is semantically dependent and grammatically subservient. Of the two constituents of compound verbs in Bangla, the vector is inflected for tense, mood, aspect, and person, and it indicates the orientation or manner of the action or process expressed by the other. Contrary to what Hook (1974) claimed, he says that, in a compound verb, the vector verb plays an important role in the selection of complements, since on occasions it motivates or induces the selection of a particular complement in place of the one that would have been selected by the main verb. In such cases phrase can be viewed as a complement of the compound verb and not just of the main verb.

The claim by Butt and Lahiri (qtd. in Butt 2010) differs from one of the commonly held views that light verbs are the result of the semantic bleaching of the main verb. There is one "underspecified entry" proposed, which gets used as both a light verb and a full verb, making a compound verb formation an instance of "co-predication where both the verbs combine to provide a single predicational head."

According to Abbi and Gopalakrishnan (1991) explicator compound verbs (EVCs) are sequences of two verbal forms, of which, the first is in stem or some nonfinite form, while the second is the morphologically finite verb marked for relevant grammatical features. Constructions consisting of a main verb and an explicator form a "complex lexeme". This forms a single unit represented by V, while constructions consisting of a main verb and auxiliary verb form a VP.

This study mainly focuses on the pattern of co-occurrence of the polar and vector verbs, the frequency of their co-occurrence, and

the resulting permissible combinations that can be identified from a POS-annotated corpora. Recognising the various ways in which CVs are defined is only a step towards understanding its structure and features, across different languages and different theories.

5. Some Challenges in Finding CVs from a POS-tagged Corpus of Gujarati

This section discusses the challenges faced in identifying CVs from a POS tagged corpus. These challenges result from the unique structure of the language and the corpus-based method employed for the study. Possible solutions to these challenges are also provided.

5.1 Unpredictability of Combinations

Not every V1 can occur with the same verbaliser/vector verb/V2 making the combinatorial patterns of V+V unpredictable. In the following examples, while the non-finite forms of V - *boli:* 'speak' and *nikali:* 'leave' combine with the V2 *padvuM* 'to fall' to form valid CVs rendering the sense of a sudden act marked by non-intentionality, the V1 *ka:dhi:* 'remove' cannot combine with the V2 *padvuM* 'to fall' to form a compound verb. On the other hand, the non-finite forms of V - *boli:* 'speak' and *ka:dhi:* 'remove' combine with the V2 na:khvum 'to throw' to form valid CVs rendering the sense of completion of an act marked with intentionality and vague sense of aggression, the V1 *nikali:* 'leave' cannot combine with the V2 na:khvum 'to throw' to form a compound verb.

(6)	boli: padvuM	boli: na:khvum
	speak+i to-fall	speak+i to-throw
	'to speak up'	'to speak out'
(7)	nikali: padvuM	*nikali: na:khvum
	leave+i to-fall	speak+i to-throw
	'to walk off'	
(8)	* ka:dhi: padvuM	ka:dhi: na:khvuM
	remove+i to-fall	remove+i to-throw
	'to dispose off'	

Given the various possibilities and restrictions of V+V patterning, the challenge then is to identify CV structures automatically, from the given POS tagged corpus, and in the process avoid other verbal sequences. Mukhopadhayay et al. (2012) have attempted to automatically identify CVs from a Bangla POS corpus, on the basis of three conditions:

1. The sequence of Verb (V1) + Verb (V2).
2. V1 ends with an inflection /-e/ (excluding ‑te).
3. V2 is a marked vector.

Marking the vectors in the language is a pre-requisite for these conditions to apply and the conditions only work well on a trained corpus. These conditions, however, fail to handle all the situations well, particularly ones with case syncretism.

In Gujarati, the CVs are composed of the non-finite form of a main verb - V1, followed by the inflected form of the vector verb- V2 as seen in the examples (9) ‑ (11) below:

(9) sukai ja:y

dry-NF go-PRS

'dries up'

(10) todi: nakhya:

break_NF throw-PST-PL

'broke off'

(11) cha:li: ni:kalyo

walk_NF leave-PST-MS

'walked off'

Following Mukhopadhayay et al. (2012), the tentative conditions for identifying Gujarati CVs have been laid down, which attempt to handle the unpredictability of Verb combinations. These conditions are:

1. Verb (V1) + Verb (V2).
2. V1 ends with non-finite form - /i:/

3. V2 is a marked vector.

In this case too, marking the vectors in the language would be a pre-requisite for these conditions to apply.

5.2 Scrambling and Inversion

One of the challenges in finding CVs from a POS tagged corpus is that of scrambling. According to Butt (1993), in Urdu, any predicate can scramble only as a unit, but not in any other order. This can also be said for Gujarati, as seen in example (12), where the verb and the auxiliary *lakhyo chhe* write-Prf.M.sg be-Pres.3P.sg 'has written' can only scramble together as a unit. She also claims that the two verbs of the CV cannot scramble away from one another; that the complex predicate can only be scrambled as a unit and that when the verbs in a complex predicate are scrambled away from one another, the result is ill-formed. This is too can be said to be the case in Gujarati, as seen in example (13) where *lakhi: lidho* write-nf take-Prf. M. Sg 'wrote up' scrambles as a unit.

(12) Mohan has written a reply.

a. Mohane uttar lakhyo chhe
   Mohan-M.erg reply-M.Acc write-Prf.M.sg be-Pres.3P.sg

b. Mohane lakhyo chhe uttar
   Mohan-M.erg write-Prf.M.sg be-Pres.3P.sg reply-M.Acc

c. * Mohane lakhyo uttar chhe
   Mohan-M.erg write-Prf.M.sg reply-M.Acc be-Pres.3P.sg

d. * Mohane chhe lakhyo uttar
   Mohan-M.erg be-Pres.3P.sg write-Prf.M.sg reply-M.Acc

e. *Mohane uttar chhe lakhyo
   Mohan-M.erg reply-M.Acc be-Pres.3P.sg write-Prf.M.sg

(13) Mohan wrote (out) a reply.

a. Mohane uttar [lakhi lidho]
   Mohan-M.erg reply-M.Acc write-nf take-Prf. M. Sg

b. Mohane [lakhi lidho] uttar
   Mohan-M.erg write-nf take-Prf. M. Sg reply-M.Acc

c. * Mohane lakhi uttar lidho
   Mohan-M.erg write-nf reply-M.Acc take-Prf. M. Sg

d. * Mohane lidho uttar lakhi
   Mohan-M.erg take-Prf. M. Sg reply-M.Acc write-nf

e. *uttar lakhi Mohane lidho

reply-M.Acc Mohan-M.erg write-nf take-Prf. M. Sg

However, in the Gujarati corpus there are occurrences of V1 and V2 inversion. Generally, in a CV, the V1 occurs in the non-finite form, but there are instances in the corpus of the V2 taking the non-finite form, and the V1 taking the finite form as seen in examples (14) and (15).

(14)TakhunI \N_NNP baane\N_NN saamethI\N_NST **aavta\V_VM joI\V_VM** bhaasaaheb\N_NNP bolya\V_VM

Takhu-M.Gen-F.Sg mother-F.Gen-F.Sg front-ABL **come-IPFV see-nf** Bhasaheb-M.NOM say-PST

‘ Seeing Takhu’s mother coming from the front Bhasaheb s’

(15) eNe\PR_PRP vidyane\N_NN akaashmaM\N_NN udti\ **V_VM** joi**V_VM** potana\PR_PRF maMtrabaLe\N_NN ene\PR_PRP utarI\V_VM

S/he-ERG vidya-ACC sky-LOC **fly-IPFV-F.SG see-nf** own-GEN.M.SG spell-binding-power-INST her-ACC land-PST

‘Seeing Vidya flying in the sky, (S/he) brought her down with his/her own spell-binding power.’

Even though such a V+V sequence does not constitute a CV and rather functions as a present participle marking temporal information, it affects the identification of the CVs in the extraction process, giving False Positives, which would further affect the frequency based count of CVs.

5.3 Insertion

Paul (2004) illustrates that the bond between the components in CV constructions differs from language to language and that the member-verbs in Hindi-Urdu and Marathi are more tightly knitted in syntax, than those of the member verbs in Bangla. In Bangla certain words can intervene between the different components, rendering the sequence discontinuous which indicates that the two Vs in a compound do not form a close-knit constituent structure. On the one hand they behave as an independent constituent on the surface, and on the other hand, they act as a single unit and provide the base for various morpho-syntactic operations. Such dual structuring can also be seen in Gujarati CVs. The occurrence of POS elements such as particles, within the CV poses a challenge in identifying the CV as a single structure, as evident in example (16).

(16) madad\N_NN karine\V_VM thaki\V_VM jauM\V_VAUX tyare\N_NST tyaM\N_NST unghi**V_VM** paN**RP_RPD** jato**V_VAUX** hato\V_VAUX

help do-PSTPTCP tire-nf go-PRS.1SG then there sleep-nf TOP go-PST.M.SG be-PST-IPVF

'When (I) was tired from helping out, I would even go to sleep there.'

However, in a set of 2845 sentences of sample corpus, only 59 sentences had insertion of POS elements in the CVs, accounting for only 2.07% of the sample. Of these 59 sentences there were 21 occurrences with an interposed negation particle and 38 occurrences with interposed inclusive, exclusive and topicalising particles. Given the low occurrence of the CV internal elements and the fact that these are generally identifiable particles; such CV structures have to be handled separately.

5.4 CVs as Inflected Word Forms

In keeping with the tentative conditions for identifying Gujarati CVs as given by Mukhopadhayay et al. (2012), marking the vectors in the language would be a pre-requisite for these conditions to apply. However, the number of vector verbs identified for listing is considerably low given that they are extracted from running texts, wherein they are inflected and occur as different instances in the frequency count. This results in their frequency getting distributed across various word forms, leading to low frequency counts of the individual CVs. The possible solution to this would involve considering a basic stemming algorithm.

5.5 Relevance of Extracted Data

Not all V+V combinations extracted from the data are relevant to the study, as seen in example (17) below.

(17) padyo\V_VM padyo\V_VM j\RP_RPD karmai:\V_VM gayo\V_VAUX .\RD_PUNC

fall-pst-MS fall-pst-MS prt-emph wither-nf go-pst-MS

'Lying around, he withered away'

A tag-based pattern matching for V_VM + V_VM on 2845 sentences returned 2773 sentences containing 4192 CVs. Considering the V+V combinations within these sentences as potential CVs, the search was refined using the pattern matching heuristic of non-finite V1. This returned 559 CVs that are then manually checked for false positives. In a larger corpus this accounts to a much larger set of CVs that need to be manually checked and identified as true CVs.

The main purpose of the study of Gujarati CVs is to identify the morpho-syntactic or semantic properties of the V1 that allows them to combine with certain light verbs but not with others, and to formulate general rules or constraints to account for the

restrictions on these combinations. However, the current paper discusses some of the challenges faced in finding/identifying CVs in a corpus-based approach.

Works Cited

Abbi, Anvita, and Devi Gopalakrishnan. "Semantics of Explicator Compound Verbs in South Asian Languages." *Language Sciences*, vol. 13, no. 2, 1991, pp. 161-180.

Butt, Miriam. *The Light Verb Jungle: Still Hacking Away.* Complex Predicates in Cross-linguistic Perspective, edited by Mengistu Amberber, Brett Baker and Mark Harvey, Cambridge UP, 2010.

Butt, Miriam, et al. "Identifying Urdu Complex Predication via Bigram Extraction." *International Conference on Computational Linguistics*, 2012.

Chakrabarti, Debasri, et al. "Hindi Compound Verbs and their Automatic Extraction." *Coling, Companion Volume: Posters*, 2008, pp. 27-30.

Dasgupta, Probal. "The Internal Grammar of Compound Verbs in Bangla." *Indian Linguistics*, vol. 38, no.3, 1977, pp. 68-85.

Das, Pradeep Kumar. *Hindi-Urdu, Grammatical Agreement and its Major Varieties.* Lincom: Europa, 2006.

Hook, Peter Edwin. "The Compound Verb in Gujarati and its Use in Connected Text." *Consciousness Manifest: Studies in Jaina Art and Iconography and Allied Subjects in Honour of Dr. U. P. Shah*, edited by R. T. Vyas, Oriental Institute of Vadodara, 1995, pp. 339-356.

Hook, Peter Edwin. *The Compound Verb in Hindi*. University of Michigan, 1974.

Mukhopadhayay, Sibanshu, et al. "Automatic Extraction of Compound Verbs from Bangla Corpora." *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, 2012, pp. 153-162.

Paul, Soma. "Composition of Compound Verbs in Bangla.". *Proceedings of the Workshop on Multi-Verb Constructions*, edited by Dorothee Beermann and Lars Hellan, Norwegian University of Science and Technology, 2003.

Paul, Soma. *An HPSG Account of Bangla Compound Verbs with LKB Implementation*. 2004. University of Hyderabad, PhD dissertation.