Revisiting VMWEs in Hindi: Annotating Layers of Predication

Kanishka Jain **Ashwini Vaidya**

Indian Institute of Technology Delhi

Goal of the Study

- Verbal Multiword Expressions (VMWEs) are a combination of verb with other lexical item(s).
- In Hindi VMWEs can be formed **morphologically** or **lexically**.
- **Hindi** not only uses a variety of VMWEs but also employs different combinatorial strategies to create new types of MWEs.
- We annotate these new categories and also refine the existing PARSEME corpus (version 1.3) by identifying key problem areas.

VMWEs in Hindi

- Hindi PARSEME corpus 1.3 [Savary et al. 2023] identifies 3 types of VMWEs:

VMWE	Preverbal	Light Verb	Causative	Gloss
VID	ləgam (rein)	ləg-a-na (put)		to control
MVC	pəɽʰ (read)	le-na (take)		to read completely
LVC.full	cori (theft)	kər-na (do)		to steal
LVC.cause	cori (theft)	kər- (do)	-va-na	to cause to steal

Table 1. Existing PARSEME VMWE Categories

- A total 1034 VMWEs have been annotated for Hindi as part of PARSEME shared task [Ramisch et al. 2020].

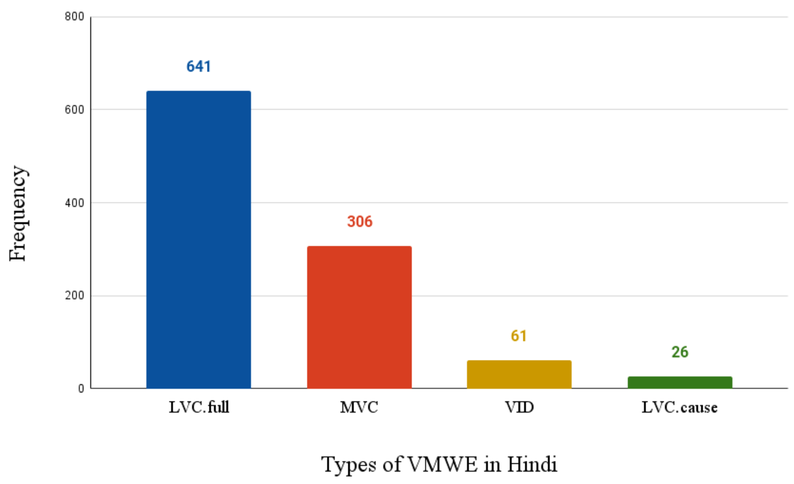

Figure 1. Number of VMWEs annotated in existing PARSEME Hindi Corpus

Morphological Causative

- Hindi verbs can be causativized using two types of morphemes: **direct and indirect**.
- **Direct causatives** are formed by attaching /-a/ morpheme or a 'null' (phonological change).
- **Indirect causatives** are formed by attching /-va/.
- VMWEs crucially **change the valency** of the verbs [Butt and King 2006].
- Similarly, causatives also change the argument structure.
- In Table 2, the base form of the verbs are monovalent, the direct causative forms are divalent, and the indirect causative forms are trivalent.

base	Direct Causative		Indirect Caus-
	Null	/a/	ative /-va/
bəT (divide)	bə̃T (to divide)		bəT-va (to cause to divide)
ubəl (boil)	ubal (to boil)		ubəl-va (to cause to boil)
kəT (cut)		kaT (to cut)	kəT-va (to cause to cut)

Table 2. Verbal paradigm for different verbs in Hindi-Urdu

- Adding morphological causatives will give us a comprehensive picture of VMWEs in the language.
- PARSEME's existing annotation schema already annotates LVC.cause distinguishing them from their non-causative counterparts.

```
(a) करवाने   करवा VERB   VM   Number=Sing|VerbForm=Inf|Cause=Yes
(b) करवा     करवा VERB   VM   Number=Sing|Person=3|Cause=Yes
```

Figure 2. Feature structure for Hindi causative verb inflected for agreement /kərvane/ in (a) and /kərva/ in (b) with the 'cause' morphological feature. Note that the lemma form form for both the verbs is /kərva/.

Stacked VMWEs

- VMWEs in Hindi are also formed by stacking two or more VMWEs to describe a single event [Butt, King, and Maxwell III 2003].
- So far, stacked VMWEs have not been implemented in any annotated corpus.
- PARSEME Hindi Corpus edition 1.3 can capture them but they are not discussed explicitly.

```
चोरी    चोरी    NOUN   NN     1:LVC.full
कर      कर      VERB   VM     1;2:MVC
डाली    डाली    AUX    VAUX   2
```

Figure 3. An example of LVC and MVC Stacked VMWE. Noun /cori/ 'theft' combines with verb /kər/ 'to do' and light verb /dalna/ 'to put'.

```
दर्शन   दर्शन   NOUN   NN     1:LVC.cause
करवा    करवा    VERB   VM     1;2:MVC
लिए     ले      AUX    VAUX   2
```

Figure 4. An example of LVC, Causative, and MVC Stacked VMWE. Noun /dərʃan/ 'sight' combines with causative verb /kərva/ 'to cause to do' and light verb /lena/ 'to take'.

Updated Annotations

- Reannotation of the corpus was done in multiple stages and varied according to each category.
- VID being the most diverse category were annotated manually using PARSEME guidelines [Ramisch et al. 2020].
- Automatic Annotation of other VMWEs exploited morphological description of tokens from **Universal Dependency** (UD) framework.
- Causatives are morphologically formed and are annotated in the feature structure of verbs as **Cause=Yes**.
- To annotate LVCs, we have exploited UD's **compound** relation and to distinguish between LVC.full and LVC.cause we have used *cause* feature.
- Annotation of MVCs was challenging and therefore we have applied a number of rules to identify and tag these verb+verb combinations.
- Since Stacked VMWEs shows recursive use of different types of MWEs, they are easily retrieved using existing annotations for LVCs, MVCs, and causatives.

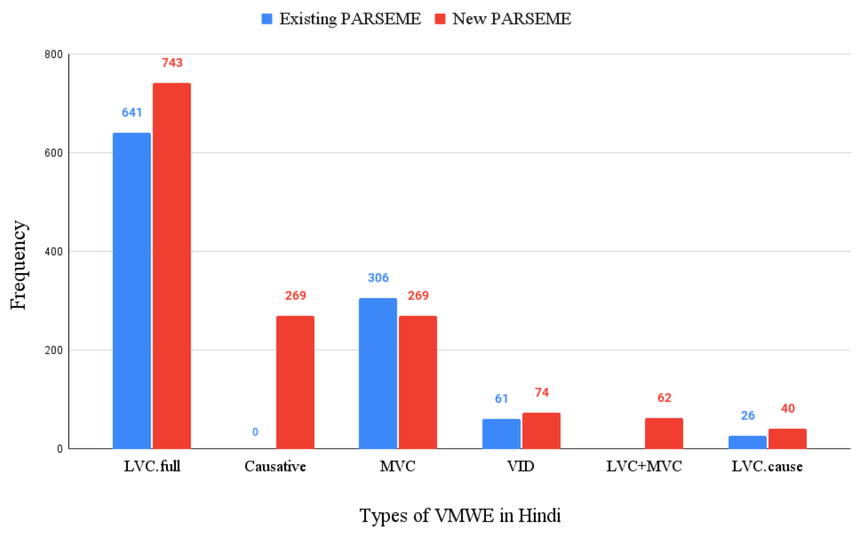

Figure 5. A comparison VMWEs in existing and updated PARSEME corpus

Existing PARSEME Hindi Corpus

We refined the annotations for VMWEs in the PARSEME Hindi Corpus edition 1.3.

- **VIDs** were often mis-annotated as a different VMWE category or an expression from another category was annotated as VID.
- The *verb + light verb* pattern is common to both Hindi MVCs as well as modals and passives.
- We made these distinct in our annotations.
  - Modals like /pa/ 'able', and /sək/ 'can or may' place an event into possible world semantics [Butt 2010] – in contrast, MVCs describe a single event.

(1) ləɽka kɪtab **pəɽʰ pa-ya**
boy.3.SG.M book.SG.F read can-PST.PERF.SG.M
'The boy could read the book.'

  - Hindi Passives appear by combining any main verb with an auxiliary verb /ja/ 'go', however they fail the tests in PARSEME shared task guidelines.
  - Preverbal in MVC should be in its base form which is not the case in passives.

(2) ləɽke-se kɪtab **pəɽʰi** **gə-yi**
boy.3.SG.M-INST book.SG.F read.PST.SG.F go-PST.PERF.SG.F
'The book was read by the boy.'

Discussion

- Our results show that VMWEs in Hindi are both productive and challenging with respect to their formation aspect.
- We also see that some VMWEs like LVC.full are more common whereas others like stacked VMWEs are rarer.
- Our study of other corpora also shows a similarity to this general distribution of VMWEs in Hindi.

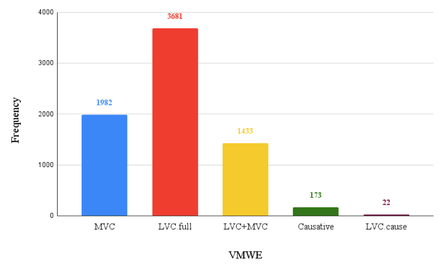

Figure 6. Distribution of VMWEs in Hindi TimeBank (∼2.1m tokens) [Goel et al. 2020]

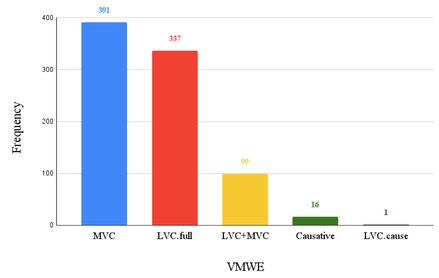

Figure 7. Distribution of VMWEs in Hindi Dialogue Corpus (∼35k tokens) [Pareek et al. 2023]

References

Pareek, B. et al. (2023). "The IIT Delhi Dialogue Corpus for Hindi". In: In Preparation.

Savary, A. et al. (2023). "PARSEME corpus release 1.3". In: *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 24–35.

Goel, P. et al. (May 2020). "Hindi TimeBank: An ISO-TimeML Annotated Reference Corpus". English. In: *16th Joint ACL - ISO Workshop on Interoperable Semantic Annotation PROCEEDINGS*. Marseille: European Language Resources Association, pp. 13–21. ISBN: 979-10-95546-48-1. URL: https://aclanthology.org/2020.isa-1.2.

Ramisch, C. et al. (2020). "Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions". In: *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*. online: Association for Computational Linguistics, pp. 107–118.

Butt, M. (2010). "The light verb jungle: Still hacking away". In: *Complex predicates in cross-linguistic perspective*, pp. 48–78.

Butt, M. and T. H. King (2006). "Restriction for morphological valency alternations : The Urdu causative". In: *Intelligent Linguistic Architectures : Variations on Themes by Ronald M. Kaplan*. Ed. by M. Butt. CSLI lecture notes 179. Stanford, California: CSLI Publications, pp. 235–258. ISBN: 1-57586-532-7.

Butt, M., T. H. King, and J. T. Maxwell III (2003). "Complex predicates via restriction". In: *Proceedings of the LFG03 Conference*,

kanishka@hss.iitd.ac.in, avaidya@iitd.ac.in