Revisiting VMWEs in Hindi: Annotating Layers of Predication

Kanishka Jain, Ashwini Vaidya
Department of Humanities and Social Sciences
Indian Institute of Technology, Delhi
{huz218481, avaidya}@iitd.ac.in

Abstract

Multiword expressions in languages like Hindi are both productive and challenging. Hindi not only uses a variety of verbal multiword expressions (VMWEs) but also employs different combinatorial strategies to create new types of multiword expressions. In this paper we are investigating two such strategies that are quite common in the language. Firstly, we describe that VMWEs in Hindi are not just lexical but also morphological. Causatives are formed morphologically in Hindi. Second, we examine Stacked VMWEs i.e. when at least two VMWEs occur together. We suggest that the existing PARSEME annotation framework can be extended to these two phenomena without changing the existing guidelines. We also propose rule-based heuristics using existing Universal Dependency annotations to automatically identify and annotate some of the VMWEs in the language. The goal of this paper is to refine the existing PARSEME corpus of Hindi for VMWEs while expanding its scope giving a more comprehensive picture of VMWEs in Hindi.

Keywords: Annotation, Stacked VMWE, Morphological Causative

1. Introduction

Verbal multiword expressions are linguistic constructions that involve multiple verbs or a combination of verb and other lexical item(s). These expressions combine to form new meanings (Baldwin and Kim, 2010). However, the non-compositional nature of these multiword expressions pose a challenge to any kind of natural language processing (NLP) task. Therefore, they have been part of multiple annotation efforts across languages.

The PARSEME shared task (Ramisch et al., 2020, 2018) is one such effort that aims to identify and annotate different types of VMWEs in multiple languages. We examine the Hindi corpus from the PARSEME shared task (Ramisch et al., 2020). In this paper, we have conducted a detailed survey of the corpus and identified some problems. A prominent issue that was prevalent across all annotation categories was missing annotations for a number of expressions. Another repeated issue that we observed is the annotation of modal constructions as multi-verbal constructions (MVCs) as both are structurally similar to each other. We address these and other issues in the existing corpus and refine the annotations to create a better quality dataset[1].

Multiword expressions in some languages are highly frequent. Hindi, for instance, in comparison to languages like English, is known to have a greater proportion of VMWEs compared to simple verbs (Vaidya et al., 2016). This productive usage of multiword expressions in the language has been captured in the PARSEME corpus edition 1.3 (Savary et al., 2023). But two additional and quite common phenomena need to be addressed. In Hindi, verbal complex allows for recursive combinations of light verb, multi-verb, and causative verbs. Sometimes all three can combine together. When two VMWEs appear together to create a single predicate then we refer to such predicate as Stacked VMWE. Further, VMWEs in Hindi are formed not only lexically (i.e. combining two or more lexical items) but also morphologically (i.e. combining two or more morphemes). In Hindi, morphological VMWEs occur as Causatives. Both, stacked and causative VMWEs are extensively used in the language but have not been explicitly annotated as such within existing annotation frameworks of the language.

The aim of this paper is twofold. First, to refine the existing corpus by addressing various issues and second, to extend its scope.

The paper is organized as follows. In Section 2 we describe different types of VMWEs found in Hindi. We also describe causatives and stacked VMWEs. Section 3 discusses the issues found in the annotations and how they have been addressed in the present study. Results and conclusion are presented in Section 4.

2. VMWEs in Hindi

2.1. PARSEME VMWEs

The PARSEME framework (Ramisch et al., 2020, 2018) has five categories of verbal multiword expressions (VMWEs) out of which three are tagged for Hindi i.e. Light Verb construction (LVC) as LVC.full and LVC.cause, Multi-Verb Construction (MVC), and Verbal Idiom (VID). The fundamental

[1]The dataset can be accessed from https://gitlab.com/kjain93/mwe_ud_hindi

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@LREC-COLING-2024, pages 98–105. May 25, 2024.

difference among these categories lie in terms of their predication strategy. A VID has at least two elements combining – a main verb and its dependent which is not restricted to any one particular lexical category as shown in (1). On the contrary, LVC and MVC are formed with a preverbal element and a light verb. The only difference between the two categories is that the preverbal element in an LVC is noun whereas in case of an MVC it is a verb as shown in (2) and (3), respectively.

(1) bəɽti mehengai pər **ləgam**
increasing.F price-hike.F on rein.SG.F
ləgana zəruri hɛ
put.INF important.F be.PRS
'It is important to control the price-hike (or inflation).'

(2) ləɽke-ne gehnõ-ki
boy.3.SG.M-ERG jewellery.PL.M-GEN.F
cori ki
theft.F do.PST.F
'The boy has stolen the jewellery.'

(3) ləɽke-ne kıtab **pəɽ^{h}**
boy.3.SG.M-ERG book.SG.F read
li
take.PST.SG.F
'The boy read the book (completely).'

Further, as mentioned above LVCs have been distinguished as LVC.full and LVC.cause. The difference is made in terms of the type of light verb used. If the light verb is 'causative' such that the subject is the cause of an event then it has been annotated as LVC.cause else as a LVC.full. An example is shown in (4). Compare it with its non-causative counterpart in (2). The subject /ləɽka/ 'boy' is the cause of an event of theft in (4) but an agent in (2). The causative meaning is expressed by the /-va/ morpheme on the verb in (4).

(4) ləɽke-ne naukar-se
boy.3.SG.M-ERG servant.3.SG.M-INST
gehnõ-ki **cori**
jewellery.PL.M-GEN.F theft.F
kər-va-yi
do-ICAUS-PST.PERF.SG.F
'The boy made the servant steal the jewellery.'

In the existing PARSEME corpus of Hindi a total of 1034 VMWEs have been annotated out of 35430 tokens as shown in Table 1. Further, it is to be noted that the frequency of VMWEs when compared to other Indo-European languages is quite high. These number are compiled from PARSEME shared tasks 2020[2] and 2018[3].

While the existing PARSEME framework covers all the prominent categories of VMWEs in Hindi, there are additional phenomena that are not present. The rest of the paper discusses two such phenomena – stacked VMWEs and causatives.

2.2. Morphological Causative

Causatives are common across natural languages. This is especially true for South-Asian languages like Hindi where any verb, theoretically, can undergo the morphological process and form causative. For instance, in (5b) the causative marker /-va/ attaches to the transitive verb /bənana/ 'build' and forms causative /bənvana/. The causativization of the transitive verb in (5a) increases the valency from two to three.

(5) a. ləɽke-ne ghər
boy.3.SG.M-ERG house-3.M
bənaya
build.PST.PERF.SG.M
'The boy built a house.'

b. ləɽke-ne bəcci-se
boy.3.SG.M-ERG girl.3.SG.F-INST
ghar
house.3.M
bən-va-ya
build-ICAUS-PST.PERF.SG.M
'The boy made the girl build the house.'

Apart from causativizing a simple verb, the language also allows causativization of light verbs[4] as shown in (4) where the light verb /ki/ 'do' is a causative.

Valency change is a property that is common to LVCs, MVCs and morphological causatives (Butt and King, 2006; Butt et al., 2008; Butt, 2010). For instance in (6a) simple verb /katna/ 'cut' has two argument positions – the servant and the tree. But in (6b) when katna/ combines with the light verb /dena/ 'give', forming an MVC, it has three argument positions. The new argument position for /laɽka/ 'boy' is licensed by the light verb /dena/ (Butt, 2010).

[2] http://multiword.sourceforge.net/mwelex2020

[3] http://multiword.sourceforge.net/lawmwecxg2018

[4] According to (Butt et al., 2008), Hindi also allows for causatives in MVC construction but we did not find examples of this in the current corpus

Langage	Tokens	VID	LVC.full	LVC.cause	MVC	Others	Total
English	124203	139	244	43	4	402	832
French	525992	2156	1878	97	22	1501	5654
German	173562	1437	311	33	0	2260	4041
Hindi	35430	61	641	26	306	0	1034
Italian	430789	1484	734	174	33	1785	4210

Table 1: Number of VMWEs in different Indo-European languages including Hindi in PARSEME shared tasks.

(6) a. naukər-ne paud^{h}a
servant.3.SG.M-ERG plant.SG.M
kata
cut.PST.PL.M
'The servant cut the plant.'

b. laɽke-ne naukər-ko
boy.3.SG.M-ERG servant.3.SG.M-DAT
paud^{h}a **katne**
plant.SG.M cut.INF.SG
dı-ya
give-PST.PERF.SG.M
'The boy let the servant cut the plant.'

This valency change is similar to causatives in example (5) where /-va/ morpheme combines with verb and license a new argument position for the causer 'girl'. This provides evidence that morphological VMWEs are similar to lexical VMWEs in Hindi. Hence, we propose to include them in the PARSEME framework.

PARSEME's existing annotation schema already annotates example like (4) as LVC.cause distinguishing them from their non-causative counterpart as in example (2) annotated as LVC.full. The addition of other causatives will then give a comprehensive picture of VMWEs in this language.

The examples discussed so far captures only one kind of causatives i.e. a causative formed by attaching /-va/ morpheme. They are also known as 'indirect causatives'. However, Hindi also has direct causatives that are formed by causativization of intransitive verbs as exemplified in (7).

(7) a. ləkɽi **jəli**
wood.SG.F burn.PST.F
'Wood burnt.'

b. ləɽke-ne ləkɽi
boy.3.SG.M-ERG wood.SG.F
jəl-a-yi
burn-DCAUS-PST.PERF.SG.F
'The boy burnt the wood.'

In (7a), the verb /jəlna/ 'burn' in intransitive whereas in (7b) the direct causative marker /-a/ is attached to the verb and forms the causative /jəlana/. Direct causatives, similar to indirect causatives, change the valency of the base verb from single argument place to two argument places. Therefore, direct causatives are also an example of morphologically formed multiword expressions.

In Hindi, direct causatives for some verbs are realized by a change in the phonological realization of the root of the verb as in (8) where the verb /d^{h}ul/ 'wash' changes to causative /d^{h}o/.

(8) a. kəpɽe **d^{h}ule**
cloth.PL wash.PERF.PL.M
'Clothes are washed.'

b. ləɽke-ne kəpɽe
boy.3.SG.M-ERG cloth.PL.M
d^{h}o-ye
wash.DCAUS-PERF.PL.M
'The boy has washed the clothes.'

These examples show that the system of morphological predication in the language is quite robust and complex. It is, therefore, essential to capture these various kinds of morphological multiword expressions to understand the representation of different types of VMWEs in Hindi. Hence, in this work we propose to annotate causatives using a morphological feature 'Cause' on verbs (see Section 3). The feature 'Cause' can effectively differentiates between the causative and non-causative forms of the verbs.

2.3. Recursive VMWEs

VMWEs in Hindi are not limited to combining two lexical items or morphological items but due to their recursive nature allow two or more VMWEs to stack describing a single event (Butt et al., 2003). An example is shown in (9) where an MVC is stacked on an LVC and results in a Stacked VMWE.

(9) ləɽke-ne gehnõ-ki
boy.3.SG.M-ERG jewellery.PL.M-GEN.F
cori kər dali
theft.F do put.PST.F
'The boy has stolen away the jewellery.'

In (9) there are three elements unlike the common pattern observed in LVCs and MVCs of predicating two elements. There is a noun /cori/ 'theft' and two verbs kər 'do' as well as /dali/ 'put'. The first or main verb can be in its base form or infinitive form whereas the second light verb is inflected for tense and aspect similar to MVC in the language.

Forming stacked VMWEs via recursion has not been implemented in an annotated corpus. Although PARSEME Hindi Corpus edition 1.3 does capture some of the stacked VMWEs as illustrated in Figure 1, it has not been discussed explicitly.

दर्शन	दर्शन	NOUN	NN	1:LVC.full
कर	कर	VERB	VM	1;2:MVC
लिए	ले	AUX	VAUX	2

Figure 1: An eaxample of LVC and MVC Stacked VMWEs in PARSEME Hindi Corpus edition 1.3. The noun /dərʃan/ 'sight' combines with the verb /kərna/ 'do' and a light verb /lena/ 'take'.

Further, recursivity in VMWEs can be seen at various levels thus resulting in layers of predication. In our example of LVC.cause in (4), the causative is stacked with an LVC forming an LVC.cause which can be further predicated with an MVC. The stacked VMWE in (10) thus shows stacking of three VMWEs – LVC+causative+MVC.

(10) ləɽke-ne naukar-se
boy.3.SG.M-ERG servant.3.SG.M-INST
gehnõ-ki **cori**
jewellery.PL.M-GEN.F theft.F
kər-va **dali**
do-ICAUS.SG.M put.PST.F
'The boy had the servant steal away the jewellery.'

The annotation of these layers of predication is shown in Figure (2).

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

Figure 2: An eaxample of LVC, Causative, and MVC Stacked VMWEs in PARSEME Hindi Corpus edition 1.3. The noun /dərʃan/ 'sight' combines with the verb /kərna/ 'do', indirect causative marker /va/, and a light verb /lena/ 'take'.

While VMWEs are formed via recursivity of existing multiword expressions, we do not intend to annotate them with a new label. Rather, we extract them using existing annotations which will be more efficient (see Section 3.2.3).

3. Enhancing the Annotations

The task of identifying multiword expressions is challenging and requires linguistic expertise. While the annotation guidelines developed as part of PARSEME shared task (Ramisch et al., 2020, 2018) standardizes the process of identification of VMWEs for many languages but there still exist various problems. In the following sections, we discuss some of the issues found in the PARSEME Hindi corpus edition 1.3 pertaining to existing annotation of VMWEs in Hindi and their refinement. We also discuss the annotations of morphological feature for causatives (Section 3.1) and representation of Stacked VMWEs (Section 3.2.3) in the existing annotation schema.

The PARSEME corpus of Hindi uses a treebank which is annotated using UD framework and therefore we could employ annotations for morphological description of tokens for automatic tagging of VMWEs.

3.1. Semi-Automated Annotation of morphological VMWEs

Beginning with causatives, we propose to add them as a morphological feature. If a verb is present in its causative form then we add 'Cause=Yes' as a boolean feature as illustrated in Figure (3). We note that Universal Dependencies guidelines have a similar feature 'Voice=Cau'[5]. In a future version of our corpus, we plan to update this feature to be in accordance with UD guidelines.

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

Figure 3: 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/.

The annotation process of causative verbs is semi-automatic as indirect causatives and one type of direct causative can be tagged using rule-based heuristics. The lemma form for /-va/ causatives have /-va/ attached however there are some discrepancies in the data therefore we have used a list of morphological endings with /-va/ morpheme varying only in terms of agreement features on the tokens to retrieve all indirect causative verbs.

The annotation of direct causatives was also challenging. Beginning with the /-a/ causatives, the UD framework does identify these causatives in their lemma. However, there are two issues in using them. First, as noted in case of indirect causative

[5] https://universaldependencies.org/u/feat/Voice.html

there are some inconsistencies with the identification of lemmas in the data. Second, Hindi also have other verbs ending with vowel /a/ like /ja/ 'go', /la/ 'get', and so on that are not causatives. Hence using only lemma leads to over-generation of tokens and to avoid that we have used multiple heuristics and manual checks while annotating the /-a/ causatives.

The second issue was with other type of direct causatives (c.f. example 8) where causative formation affects the phonological realization of the root and we get irregular forms. Since there is no particular pattern which can be exploited to identify these kind of verbs we have annotated them manually. A total of 269 causatives have been annotated – 165 automatically and 104 manually.

3.2. Automated Annotations of lexical VMWEs

Annotation of LVCs and MVCs was done in two stages, that is, automatic annotation using python scripts followed by manual adjudication. After annotating LVCs and MVCs we have extracted Stacked VMWEs.

3.2.1. LVCs

In this work, we aim to comprehensively annotate all the occurrences of VMWEs in the corpus. While examining the PARSEME corpus we observed that despite passing tests from the PARSEME guidelines a number of MVWEs were not annotated. Though it was true for all the categories, it was especially seen in case of LVCs (see Table 2 for comparision). Therefore, we used the dependency relation to find all the instances of LVCs in the corpus. Particularly, the 'compound' dependency relation that already identifies these noun+verb pairs have been used as in Figure 4.

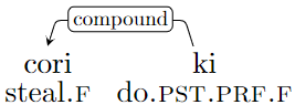

Figure 4: Compound dependency relation as tagged in UD framework for LVCs

All the missing LVCs were added to the existing corpus according to PARSEME guidelines. In order to distinguish between LVC.full and LVC.cause we use feature 'cause', annotated previously. For the purpose of this work, we have limited LVC.cause to only indirect caustives and have not included direct causatives.

We have also manually adjudicated the corpus using PARSEME tests for LVCs to remove any erroneous cases that have been annotated. Since, automatic annotations were dependent on UD dependency relation, we found few instances where nouns that were not abstract have been identified to be in compound relation with a verb as shown in (11)

(11) dʰən lı-ya
money.M take-PST.PERF.M
'took money'

In (11), /dʰən/ 'money' is annotated for compound relation with verb lıya 'take'. These were not annotated as LVCs.

Data	LVC full	LVC cause
PARSEME	641	26
New	743	40

Table 2: Number of LVCs in existing PARSEME corpus and the new corpus.

3.2.2. MVCs

MVCs as discussed in Section 1 are formed by the combination of verb with a light verb. However, this pattern is confusable with other types of constructions in Hindi. For instance, both modal and passive constructions are superficially similar to MVCs. Modal verbs include examples like /pa 'able', and sək/ 'can/may' (example 12). /pa/ is ambiguous such that the same form occurs both as a simple verb meaning 'to get' and as a ability modal (Bhatt et al., 2011). As a simple verb, it can form a complex predicate and occur as a preverbal but it does not occur as a light verb. The current guidelines of PARSEME includes it as a light verb, however according to our current analysis the guidelines for Hindi needs to be updated to prevent confusion with modals.

For both MVCs and modals, the main verb appears in its base form while light verbs and modals are inflected for agreement features (Butt and Ramchand, 2005), as shown in (12).

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

Constructions like (12) will pass the PARSEME tests for tagging MVCs, however, semantically there is a difference between light verbs and modals. Light verbs contribute sub-event information as seen in (13), where light verb /dıya/ 'give' contributes permissive meaning to the event (Butt,

1995). Modals, on the other hand, place an event into possible world semantics (Butt, 2010) (example (12)).

(13) ləɽke-ne naukar-ko
boy.3.SG.M-ERG servant.3.SG.M-DAT
xət **pəɽʰne dı-ya**
letter.SG.M read.INF give-PST.PERF.SG.M
'The boy let the servant read the letter.'

Similarly, verbs in passive constructions appear by combining any main verb with an auxiliary verb /ja/ 'go' as shown in (14). The /ja/ 'go' can participate in a number of constructions. It can be used as a simple verb with the meaning ' to go', as a light verb with the meaning 'with force' and also as an auxiliary when a sentence is passivized. On the surface, the passive resembles MVCs where two verbs are predicated and are incorrectly annotated as MVCs in the current PARSEME corpus of Hindi at several places.

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

The main verb in passives, for example pəɽʰi 'read', in (14), is inflected for tense and aspect which violates the first test of PARSEME guidelines for MVCs that the first verb (V-dep) should be non-finite. Therefore, passives clearly are not a case of VMWEs in Hindi.

Annotating MVCs was a little challenging as there is no direct relation in UD framework that can identify these verb+verb constructions. Further, we have to avoid constructions like modals and passives to be falsely tagged. Therefore, we have applied a number of rules to identify MVCs.

We have first filtered verbs that were tagged as 'VM' (main verb) for their xpos and are followed by auxiliary verbs (tagged as VAUX). Since, VAUX in all of these annotations includes any verb that has not been annotated as the main verb of the sentence, we decided to use a list of commonly used auxiliaries in Hindi including copulas, progressive marker, modals, and /vala/ to filter any false positive MVC cases, thereby also resolving the issue of modal constructions being tagged as MVCs. We have also filtered main verbs for any tense, aspect, and agreement inflections resulting in verbs that are in their base or infinitive form to avoid tagging of passives.

MVCs have also been added to the existing annotations according to the guidelines. If it already exists then we do not make any changes. It was followed by manual adjudication of the data to remove any false positive cases.

On comparing with original numbers (c.f Table 1), the total number of MVCs has dropped to 269. The reason is the removal of modals and passives from the data.

3.2.3. Stacked VMWEs

In Section 2.3 we have mentioned that we are not introducing any new label for Stacked VMWEs. As discussed, Stacked VMWEs shows recursive use of different types of multiword expressions occurring as a single predicate. Therefore, they can be easily retrieved using existing annotations for LVCs, MVCs, and causatives. For instance, as illustrated in Figure 1, we can extract by looking for verbs that are annotated for both LVCs and MVCs. Table 3 shows the frequency of stacked VMWEs. Also, note that since PARSEME has not reported the numbers for Stacked VMWEs in their previous editions of the language we have kept it as null.

Data	LVC.full +MVC	LVC.cause +MVC
PARSEME	null	null
New	61	1

Table 3: Number of Stacked VMWEs in the existing PARSEME corpus as compared to the New corpus.

The above table also highlights the fact that stacking of one VMWE onto another increases the complexity of the predicates and therefore occurs less frequently when compared to other VMWEs. As we can see that there was only one instance of LVC+causative+MVC kind of expression.

3.3. Verbal Idioms

Multiword Expressions are known for their non-compositionality with VIDs being the most diverse category such that detection of VIDs by automatic means was challenging. There were two types of issues. First, when a VID was tagged with a different VMWE category. Second, when an expression from another VMWE category was annotated as VID. Therefore, we have annotated them manually using PARSEME guidelines (Ramisch et al., 2020). These led to changes in the overall numbers of VIDs. As we can see in Table 4 the numbers have increased after the reannotation of the data especially after identifying the miscategorized VIDs.

4. Results and Conclusion

The main aim of this study was to enhance the existing PARSEME Hindi corpus by expanding its scope

Data	VID
PARSEME	61
New	74

Table 4: Number of VIDs in the existing PARSEME corpus as compared to the New corpus.

to other phenomena that results in the formation of different types of multiword expressions. Towards this goal we have proposed to annotate causatives via a morphological feature and to extract stacked VMWEs by using the existing annotations of other VMWEs. The new corpus now have the following categories – VID, LVC.full, LVC.cause, MVC, Causative, and Stacked VMWE.

Further, the results show that Hindi frequently employs VMWEs as shown in Figure 5. LVC.full are more common where as stacked VMWEs are rarer.

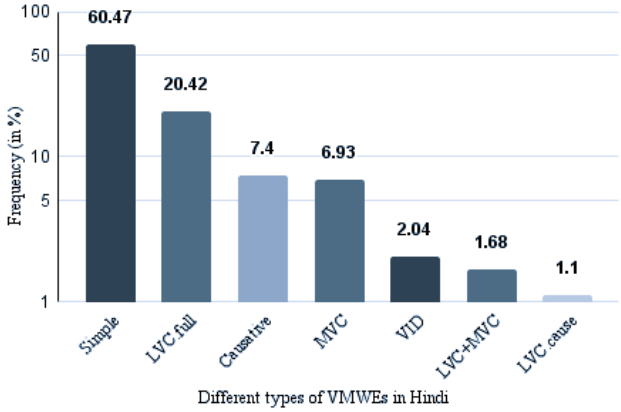

Figure 5: Frequency distribution of Hindi verbs in the new corpus.

Both Stacked VMWEs as well as causatives are infrequent as compared to other VMWE categories in all types of Hindi corpora. Our survey of corpora from other genres e.g., the Hindi TimeBank (Goel et al., 2020) and the IIT Delhi Dialogue Corpus for Hindi (Pareek et al., 2023) shows that Stacked VMWEs and causatives are consistently used (although they are relatively infrequent). We believe it is important to include these categories in the annotation framework to have a complete picture of VMWEs in Hindi.

Another goal of this study was to refine the existing annotations. For this, we have conducted a survey and identified a number of issues in the corpus. We have added annotations for the missing cases across different categories of VMWEs and removing any erroneous cases. The refinement process involved a combination of an automatic and manual annotation followed by adjudication. In case of automatic annotations we have described a method using UD framework to annotate some of the categories.

5. References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Rajesh Bhatt, Tina Bögel, Miriam Butt, Annette Hautli, and Sebastian Sulger. 2011. Urdu/hindi modals. *Proceedings of the LFG11 conference*, pages 47–67.

Miriam Butt. 1995. *The structure of complex predicates in Urdu*. Center for the Study of Language (CSLI).

Miriam Butt. 2010. The light verb jungle: Still hacking away. *Complex predicates in cross-linguistic perspective*, pages 48–78.

Miriam Butt and Tracy Holloway King. 2006. Restriction for morphological valency alternations : The Urdu causative. In Miriam Butt, editor, *Intelligent Linguistic Architectures : Variations on Themes by Ronald M. Kaplan*, number 179 in CSLI lecture notes, pages 235–258. CSLI Publications, Stanford, California.

Miriam Butt, Tracy Holloway King, and John T Maxwell III. 2003. Complex predicates via restriction. In *Proceedings of the LFG03 Conference*, pages 92–104.

Miriam Butt, Tracy Holloway King, and Gillian Ramchand. 2008. Complex predication: How did the child pinch the elephant. *Reality Exploration and Discovery: Pattern Interaction in Language & Life. A Festschrift for KP Mohanan*, pages 231–256.

Miriam Butt and Gillian Ramchand. 2005. Complex aspectual structure in Hindi/Urdu. In *The Syntax of Aspect*, pages 117–153. Oxford University Press Oxford.

Pranav Goel, Suhan Prabhu, Alok Debnath, Priyank Modi, and Manish Shrivastava. 2020. Hindi TimeBank: An ISO-TimeML annotated reference corpus. In *16th Joint ACL - ISO Workshop on Interoperable Semantic Annotation PROCEEDINGS*, pages 13–21, Marseille. European Language Resources Association.

Benu Pareek, Mudafia Zafar, Karan Yadav, Meghna Hooda, Ashwini Vaidya, and Samar Husain. 2023. The IIT Delhi Dialogue Corpus for Hindi. In Preparation.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 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*, pages 107–118, online. Association for Computational Linguistics.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Ashwini Vaidya, Sumeet Agarwal, and Martha Palmer. 2016. Linguistic features for Hindi light verb construction identification. In *International Conference on Computational Linguistics*.