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 VMWEs.
- We annotate these new categories and also refine the existing PARSEME corpus (version 1.3) by identifying key problem areas.

VMWEs in Hindi

<table>
<thead>
<tr>
<th>VMWE</th>
<th>Preverbal</th>
<th>Light Verb</th>
<th>Causative</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>VID</td>
<td>logan</td>
<td>be-ga-na</td>
<td>to control</td>
<td></td>
</tr>
<tr>
<td>MVC</td>
<td>p@ó</td>
<td>be-na</td>
<td>to read completely</td>
<td></td>
</tr>
<tr>
<td>LVC.full</td>
<td>bar-na</td>
<td>do</td>
<td>to steal</td>
<td></td>
</tr>
<tr>
<td>LVC.cause</td>
<td>bar-na</td>
<td>do</td>
<td>to cause to steal</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Existing PARSEME VMWE Categories

Figure 2. Feature structure for Hindi causative verb inflected for agreement /-va/

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 VMWEs, they are easily retrieved using existing annotations for LVCs, MVCs, and causatives.

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 attaching /-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.

<table>
<thead>
<tr>
<th>base</th>
<th>Direct Causative</th>
<th>Indirect Causative</th>
</tr>
</thead>
<tbody>
<tr>
<td>to (divide)</td>
<td>/ja/ (to divide)</td>
<td>(to cause to divide)</td>
</tr>
<tr>
<td>ubal (boil)</td>
<td>/ja/ (boil)</td>
<td>(to cause to boil)</td>
</tr>
<tr>
<td>kT (cut)</td>
<td>/ja/ (cut)</td>
<td>(to cause to cut)</td>
</tr>
</tbody>
</table>

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 annotate causative distinguishing them from their non-causative counterparts.

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.

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 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 VMCEs as well as modals and passives.
- We made these distinct in our annotations.
- Modals like /ja/ ‘able’, and /al/ ‘can or may’ place an event into possible world semantics [Bhatt 2010] – in contrast, MVCs describe a single event.

References