Light Verb Constructions in Universal Dependencies for South Asian Languages

Abishek Stephen, Daniel Zeman

May 25, 2024
The LVCs belong to the class of complex predicates with a wide range of combinatorial potential where a verb (VERB) can combine with adjectives (ADJ), adverbs (ADV) or nouns (NOUN).
Light Verb Constructions

- The LVCs belong to the class of complex predicates with a wide range of combinatorial potential where a verb (VERB) can combine with adjectives (ADJ), adverbs (ADV) or nouns (NOUN).

- The current annotations in the treebanks of many languages in Universal Dependencies (UD) treat the LVCs as combinations of lexemes that morphosyntactically behave as single words and mark them using the dependency relation compound or its subtype compound:lvc.
• The LVCs belong to the class of complex predicates with a wide range of combinatorial potential where a verb (VERB) can combine with adjectives (ADJ), adverbs (ADV) or nouns (NOUN).

• The current annotations in the treebanks of many languages in Universal Dependencies (UD) treat the LVCs as combinations of lexemes that morphosyntactically behave as single words and mark them using the dependency relation compound or its subtype compound:lvc.

• In the case of South Asian languages this is problematic given the surface-identical noun incorporations and object-verb sequences.
Examples from Hindi

pula kā nirmāṇa kiyā gayā hai
bridge of construction done gone is
‘The bridge has been constructed.’
Examples from Hindi

'The bridge has been constructed.'

(We) will work together in the field too.
Examples from Hindi

pula kā nirmāṇa kiyā gayā hai

bridge of construction done gone is

‘The bridge has been constructed.’
Examples from Hindi

- **Sentence 1:**
  - Hindi: पुला का निर्माण किया गया है।
  - English: The bridge has been constructed.

- **Sentence 2:**
  - Hindi: क्षेत्र में भी मिलकर काम करेंगे।
  - English: (We) will work together in the field too.
Examples from Telugu

'We should eat.'

'We should work.'
Examples from Telugu

We should eat.

You should work.

Structural Composition of LVCs
Morphosyntax of LVCs
LVCs in UD Revisited
Conclusion
Examples from Telugu

We should eat.

We should eat.

You should work.
Examples from Telugu

'We should eat.'

'We should work.'

You should work.'
Data

We use the treebanks of Indo-Aryan and Dravidian languages from UD (de Marneffe et al., 2021) version 2.13.

<table>
<thead>
<tr>
<th>Language</th>
<th>Treebank</th>
<th>Sentences</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanskrit</td>
<td>Vedic</td>
<td>3,997</td>
<td>27,117</td>
</tr>
<tr>
<td>Sanskrit</td>
<td>UFAL</td>
<td>230</td>
<td>1,843</td>
</tr>
<tr>
<td>Hindi</td>
<td>HDTB</td>
<td>16,649</td>
<td>351,704</td>
</tr>
<tr>
<td>Hindi</td>
<td>PUD</td>
<td>1,000</td>
<td>23,829</td>
</tr>
<tr>
<td>Urdu</td>
<td>UDTB</td>
<td>5,130</td>
<td>138,077</td>
</tr>
<tr>
<td>Kangri</td>
<td>KDTB</td>
<td>288</td>
<td>2,514</td>
</tr>
<tr>
<td>Bhojpuri</td>
<td>BHTB</td>
<td>357</td>
<td>6,665</td>
</tr>
<tr>
<td>Bengali</td>
<td>BRU</td>
<td>56</td>
<td>320</td>
</tr>
<tr>
<td>Marathi</td>
<td>UFAL</td>
<td>466</td>
<td>3,847</td>
</tr>
<tr>
<td>Sinhala</td>
<td>STB</td>
<td>100</td>
<td>880</td>
</tr>
<tr>
<td>Telugu</td>
<td>MTG</td>
<td>1,328</td>
<td>6,465</td>
</tr>
<tr>
<td>Tamil</td>
<td>TTB</td>
<td>600</td>
<td>9,581</td>
</tr>
<tr>
<td>Tamil</td>
<td>MWTT</td>
<td>534</td>
<td>2,584</td>
</tr>
<tr>
<td>Malayalam</td>
<td>UFAL</td>
<td>218</td>
<td>2,403</td>
</tr>
</tbody>
</table>
We use the treebanks of Indo-Aryan and Dravidian languages from UD (de Marneffe et al., 2021) version 2.13.

<table>
<thead>
<tr>
<th>Language</th>
<th>Treebank</th>
<th>Sentences</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanskrit</td>
<td>Vedic</td>
<td>3,997</td>
<td>27,117</td>
</tr>
<tr>
<td>Sanskrit</td>
<td>UFAL</td>
<td>230</td>
<td>1,843</td>
</tr>
<tr>
<td>Hindi</td>
<td>HDTB</td>
<td>16,649</td>
<td>351,704</td>
</tr>
<tr>
<td>Hindi</td>
<td>PUD</td>
<td>1,000</td>
<td>23,829</td>
</tr>
<tr>
<td>Urdu</td>
<td>UDTB</td>
<td>5,130</td>
<td>138,077</td>
</tr>
<tr>
<td>Kangri</td>
<td>KDTB</td>
<td>288</td>
<td>2,514</td>
</tr>
<tr>
<td>Bhojpuri</td>
<td>BHTB</td>
<td>357</td>
<td>6,665</td>
</tr>
<tr>
<td>Bengali</td>
<td>BRU</td>
<td>56</td>
<td>320</td>
</tr>
<tr>
<td>Marathi</td>
<td>UFAL</td>
<td>466</td>
<td>3,847</td>
</tr>
<tr>
<td>Sinhala</td>
<td>STB</td>
<td>100</td>
<td>880</td>
</tr>
<tr>
<td>Telugu</td>
<td>MTG</td>
<td>1,328</td>
<td>6,465</td>
</tr>
<tr>
<td>Tamil</td>
<td>TTB</td>
<td>600</td>
<td>9,581</td>
</tr>
<tr>
<td>Tamil</td>
<td>MWTT</td>
<td>534</td>
<td>2,584</td>
</tr>
<tr>
<td>Malayalam</td>
<td>UFAL</td>
<td>218</td>
<td>2,403</td>
</tr>
</tbody>
</table>
We use the treebanks of Indo-Aryan and Dravidian languages from UD (de Marneffe et al., 2021) version 2.13.

<table>
<thead>
<tr>
<th>Language</th>
<th>Treebank</th>
<th>Sentences</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanskrit</td>
<td>Vedic</td>
<td>3,997</td>
<td>27,117</td>
</tr>
<tr>
<td>Sanskrit</td>
<td>UFAL</td>
<td>230</td>
<td>1,843</td>
</tr>
<tr>
<td>Hindi</td>
<td>HDTB</td>
<td>16,649</td>
<td>351,704</td>
</tr>
<tr>
<td>Hindi</td>
<td>PUD</td>
<td>1,000</td>
<td>23,829</td>
</tr>
<tr>
<td>Urdu</td>
<td>UDTB</td>
<td>5,130</td>
<td>138,077</td>
</tr>
<tr>
<td>Kangri</td>
<td>KDTB</td>
<td>288</td>
<td>2,514</td>
</tr>
<tr>
<td>Bhojpuri</td>
<td>BHTB</td>
<td>357</td>
<td>6,665</td>
</tr>
<tr>
<td>Bengali</td>
<td>BRU</td>
<td>56</td>
<td>320</td>
</tr>
<tr>
<td>Marathi</td>
<td>UFAL</td>
<td>466</td>
<td>3,847</td>
</tr>
<tr>
<td>Sinhala</td>
<td>STB</td>
<td>100</td>
<td>880</td>
</tr>
<tr>
<td>Telugu</td>
<td>MTG</td>
<td>1,328</td>
<td>6,465</td>
</tr>
<tr>
<td>Tamil</td>
<td>TTB</td>
<td>600</td>
<td>9,581</td>
</tr>
<tr>
<td>Tamil</td>
<td>MWTT</td>
<td>534</td>
<td>2,584</td>
</tr>
<tr>
<td>Malayalam</td>
<td>UFAL</td>
<td>218</td>
<td>2,403</td>
</tr>
</tbody>
</table>
Most of these treebanks use the dependency label `compound` to mark the verbo-nominal compounds or LVCs but the Bengali, Marathi, and Sinhala treebanks use the language-specific dependency sub-type label `compound:lvc`.

---

**Figure 1:** Compound analysis in Hindi (HDTB).
Indo-Aryan Languages II

Bengali, Bhojpuri and Kangri also present a similar picture where the verbs ‘to do’ and ‘to be’ persistently head such constructions.

‘I will finish the writing and watch the cartoon.’

Figure 2: Compound analysis in Bengali (BRU).
Sinhala happens to be the only Indo-Aryan language in UD to select the noun as a head for LVCs (Liyanage et al., 2023).

‘Both of them ruined their reputation.’

**Figure 3:** A verbo-nominal compound in Sinhala (STB), headed by the nominal node.
For LVCs, only the compounds with the do-verb *ceyyuka* are labeled as `compound:lvc` in the Malayalam UFAL treebank (Stephen and Zeman, 2023).

Figure 4: A verbo-nominal compound in Malayalam (UFAL), headed by the nominal node.
The UD taxonomy has a more relaxed definition of compounds: it states that the compound relation should be used for combinations of lexemes that morphosyntactically behave as single words, and lexicalization or semantic idiomaticity should not be a criterion for identifying compounds.
Figure 5: Multiple noun-verb pairs in Hindi (HDTB)

“When challenged by the army, the terrorists started firing.”
Expressions that would qualify should have a single argument structure or in other words, the syntactic head of an LVC should select all the required arguments and the dependent noun should neither be modified nor have an argument structure of its own. But in the case of the Indo-Aryan languages, this does not seem to be the case!!

Figure 6: A verbo-nominal compound in Marathi (UFAL), arguments attached to the nominal node.
Figure 7: A verbo-nominal compound in Marathi (UFAL), arguments attached to the verbal node.
According to Haspelmath (2023a), incorporation is an event-denoting noun-verb compound construction in which the noun occupies an argument slot of the verb and occurs in a position where nominal patient arguments cannot occur.
According to Haspelmath (2023a), incorporation is an event-denoting noun-verb compound construction in which the noun occupies an argument slot of the verb and occurs in a position where nominal patient arguments cannot occur.

In most Indo-Aryan languages, verbo-nominal predicates must be analyzed as a lexical category but paradoxically enough, the noun is on par with a syntactically independent argument (Mohanan, 1995).
According to Haspelmath (2023a), incorporation is an event-denoting noun-verb compound construction in which the noun occupies an argument slot of the verb and occurs in a position where nominal patient arguments cannot occur.

In most Indo-Aryan languages, verbo-nominal predicates must be analyzed as a lexical category but paradoxically enough, the noun is on par with a syntactically independent argument (Mohanan, 1995).

Therefore, even though noun incorporation is a type of compounding of a syntactic object with the verb, both the object and the verb can have their own argument structures.
• According to Haspelmath (2023a), incorporation is an event-denoting noun-verb compound construction in which the noun occupies an argument slot of the verb and occurs in a position where nominal patient arguments cannot occur.

• In most Indo-Aryan languages, verbo-nominal predicates must be analyzed as a lexical category but paradoxically enough, the noun is on par with a syntactically independent argument (Mohanan, 1995).

• Therefore, even though noun incorporation is a type of compounding of a syntactic object with the verb, both the object and the verb can have their own argument structures.

• Incorporated nouns do not take case or plural markers and external modifiers, they are morphosyntactically different from the regular object nouns.
Case Marking

- Hindi, Urdu, and some other Indo-Aryan languages follow a split-ergative pattern. Animate direct objects use the postposition *ko*. Inanimate direct objects may omit the postposition *ko* in the accusative case.
Hindi, Urdu, and some other Indo-Aryan languages follow a split-ergative pattern. Animate direct objects use the postposition *ko*. Inanimate direct objects may omit the postposition *ko* in the accusative case.

Bhojpuri uses the same postposition (*ke*) for accusative, dative, and genitive, making it less obvious when it is selected by the nominal and not the verb.

Figure 8: A verbo-nominal compound in Bhojpuri (BHTB) where the nominal conjunct *āyojana* ‘organizing’ selects the argument *kājakarama* ‘event’ case marked using the postposition *ke* ‘.'
• Agreement with the verb in transitive-perfective clauses is another signal that the nominal of an LVC candidate is an object rather than part of a compound.
Agreement

• Agreement with the verb in transitive-perfective clauses is another signal that the nominal of an LVC candidate is an object rather than part of a compound.

• And it can also attest to the opposite: In the Hindi example, *mere pitā ne pūjā śurū kar dī hai* ‘my father has started the prayer’, the verb has a feminine form, agreeing with *pūjā*, while both *pitā* ‘father’ and *śurū* ‘start’ are masculine.
• Agreement with the verb in transitive-perfective clauses is another signal that the nominal of an LVC candidate is an object rather than part of a compound.

• And it can also attest to the opposite: In the Hindi example, *mere pitā ne pūjā śurū kar dī hai* ‘my father has started the prayer’, the verb has a feminine form, agreeing with *pūjā*, while both *pitā* ‘father’ and *śurū* ‘start’ are masculine.

• In Telugu, the verb agrees with the subject when it is in the nominative case, whereas when there is a dative “subject”, the verb agrees with the incorporated noun (Nadimpalli and Lakshmi, 2022).
Agreement

• Agreement with the verb in transitive-perfective clauses is another signal that the nominal of an LVC candidate is an object rather than part of a compound.

• And it can also attest to the opposite: In the Hindi example, \textit{mere pitā ne pūjā śurū kar dī hai} ‘my father has started the prayer’, the verb has a feminine form, agreeing with \textit{pūjā}, while both \textit{pitā ‘father’} and \textit{śurū ‘start’} are masculine.

• In Telugu, the verb agrees with the subject when it is in the nominative case, whereas when there is a dative “subject”, the verb agrees with the incorporated noun (Nadimpalli and Lakshmi, 2022).

• In many instances of noun-verb sequences agreement between the noun and the verb is observed and represents a deviation from typical compound behavior.
Modification

- One of the signs of compounds is that their parts (and especially the dependent part) cannot be modified individually.
• One of the signs of compounds is that their parts (and especially the dependent part) cannot be modified individually.

• In Kangri, the nominal *galla* ‘matter’ is modified by the determiner *isadī* ‘this’, suggesting that *galla mannī* is not a compound.

Figure 9: Compound analysis in Kangri (KDTB).
Noun-verb compounds are very frequent in the current UD treebanks of South Asian languages.
The Current State

• Noun-verb compounds are very frequent in the current UD treebanks of South Asian languages.

• In Hindi HDTB, there are 6187 such compounds with the 5 most common verbs alone (out of which 4159 occurrences belong just to karānā ‘to do’). A similar pattern is found in the smaller Urdu treebank: 3542 occurrences with the top 5 verbs, including 2346 with krnā ‘to do’.

For example, Hindi bāta karānā ‘to talk’ is a relatively frequent expression and it is usually annotated as compound (118 instances), though occasionally it is annotated as obj (25 instances).

We can conclude that in the present versions of the treebanks of South Asian languages, the treatment of noun-verb sequences or L VCs as compounds is not consistent because the interplay of surface level similarities between real noun-verb compounds and noun incorporations somehow weigh down the morphosyntactic cues.
The Current State

• Noun-verb compounds are very frequent in the current UD treebanks of South Asian languages.

• In Hindi HDTB, there are 6187 such compounds with the 5 most common verbs alone (out of which 4159 occurrences belong just to karanā ‘to do’). A similar pattern is found in the smaller Urdu treebank: 3542 occurrences with the top 5 verbs, including 2346 with krnā ‘to do’.

• For example, Hindi bāta karanā ‘to talk’ is a relatively frequent expression and it is usually annotated as compound (118 instances), though occasionally it is annotated as obj (25 instances).
The Current State

- Noun-verb compounds are very frequent in the current UD treebanks of South Asian languages.

- In Hindi HDTB, there are 6187 such compounds with the 5 most common verbs alone (out of which 4159 occurrences belong just to *karanā* ‘to do’). A similar pattern is found in the smaller Urdu treebank: 3542 occurrences with the top 5 verbs, including 2346 with *krnā* ‘to do’.

- For example, Hindi *bāta karanā* ‘to talk’ is a relatively frequent expression and it is usually annotated as *compound* (118 instances), though occasionally it is annotated as *obj* (25 instances).

- We can conclude that in the present versions of the treebanks of South Asian languages, the treatment of noun-verb sequences or LVCs as compounds is not consistent because the interplay of surface level similarities between real noun-verb compounds and noun incorporations somehow weigh down the morphosyntactic cues.
What could be done?

- There should not be a problem if noun-verb compounds satisfying the UD guidelines are marked as `compound:lvc` just to differentiate it from other type of compounds.

- This would also handle most of the noun incorporations, but once the nominal participant is case marked, modified or triggering verbal agreement, the sequence should be analyzed differently.

- One of the solutions could be to label the relation `obj:lvc`, modifying Vincze et al. (2017)’s proposal to fit the current UD version.

- By doing so, there will be a three-way distinction:
  - Noun incorporations (with a single argument structure) marked as `compound:lvc`.
  - Object-verb sequences marked as `obj`.
  - Noun-incorporations with individual noun and verb argument structures as `obj:lvc`.

- LVCs
- Cross-Lingual Comparison
- Structural Composition of LVCs
- Morphosyntax of LVCs
- LVCs in UD Revisited
- Conclusion
• There should not be a problem if noun-verb compounds satisfying the UD guidelines are marked as `compound:lvc` just to differentiate it from other type of compounds.

• This would also handle most of the noun incorporations, but once the nominal participant is case marked, modified or triggering verbal agreement, the sequence should be analyzed differently.
What could be done?

• There should not be a problem if noun-verb compounds satisfying the UD guidelines are marked as `compound:lvc` just to differentiate it from other type of compounds.

• This would also handle most of the noun incorporations, but once the nominal participant is case marked, modified or triggering verbal agreement, the sequence should be analyzed differently.

• One of the solutions could be to label the relation `obj:lvc`, modifying Vincze et al. (2017)’s proposal to fit the current UD version.
What could be done?

- There should not be a problem if noun-verb compounds satisfying the UD guidelines are marked as `compound:lvc` just to differentiate it from other type of compounds.

- This would also handle most of the noun incorporations, but once the nominal participant is case marked, modified or triggering verbal agreement, the sequence should be analyzed differently.

- One of the solutions could be to label the relation `obj:lvc`, modifying Vincze et al. (2017)’s proposal to fit the current UD version.

- By doing so, there will be a three-way distinction.
What could be done?

• There should not be a problem if noun-verb compounds satisfying the UD guidelines are marked as `compound:1vc` just to differentiate it from other type of compounds.

• This would also handle most of the noun incorporations, but once the nominal participant is case marked, modified or triggering verbal agreement, the sequence should be analyzed differently.

• One of the solutions could be to label the relation `obj:1vc`, modifying Vincze et al. (2017)’s proposal to fit the current UD version.

• By doing so, there will be a three-way distinction.
  
  • Noun incorporations (with a single argument structure) marked as `compound:1vc`. 
What could be done?

• There should not be a problem if noun-verb compounds satisfying the UD guidelines are marked as compound:lvc just to differentiate it from other type of compounds.

• This would also handle most of the noun incorporations, but once the nominal participant is case marked, modified or triggering verbal agreement, the sequence should be analyzed differently.

• One of the solutions could be to label the relation obj:lvc, modifying Vincze et al. (2017)’s proposal to fit the current UD version.

• By doing so, there will be a three-way distinction.
  
  • Noun incorporations (with a single argument structure) marked as compound:lvc.

  • Object-verb sequences marked as obj.
What could be done?

- There should not be a problem if noun-verb compounds satisfying the UD guidelines are marked as `compound:lvc` just to differentiate it from other type of compounds.

- This would also handle most of the noun incorporations, but once the nominal participant is case marked, modified or triggering verbal agreement, the sequence should be analyzed differently.

- One of the solutions could be to label the relation `obj:lvc`, modifying Vincze et al. (2017)’s proposal to fit the current UD version.

- By doing so, there will be a three-way distinction.
  
  - Noun incorporations (with a single argument structure) marked as `compound:lvc`.
  
  - Object-verb sequences marked as `obj`.
  
  - Noun-incorporations with individual noun and verb argument structures as `obj:lvc`.
We have presented morphosyntactic clues for identifying light verb constructions in South Asian languages, which could prove instrumental in achieving consistent annotations of `compound` and `compound:lvc` dependency relations.

While LVCs as semantically idiosyncratic constructions are widespread in these languages, we have shown that in many cases their syntactic behavior is transparent or very close to standard object-verb constructions. Their compound analysis should be reconsidered and the annotation could be changed to `obj` or `obj:lvc` based on the type of argument sharing.
We have presented morphosyntactic clues for identifying light verb constructions in South Asian languages, which could prove instrumental in achieving consistent annotations of `compound` and `compound:lvc` dependency relations.

While LVCs as semantically idiosyncratic constructions are widespread in these languages, we have shown that in many cases their syntactic behavior is transparent or very close to standard object-verb constructions.
• We have presented morphosyntactic clues for identifying light verb constructions in South Asian languages, which could prove instrumental in achieving consistent annotations of compound and compound:lvc dependency relations.

• While LVCs as semantically idiosyncratic constructions are widespread in these languages, we have shown that in many cases their syntactic behavior is transparent or very close to standard object-verb constructions.

• Their compound analysis should be reconsidered and the annotation could be changed to obj or obj:lvc based on the type of argument sharing.
Thank you!

stephen,zeman@ufal.mff.cuni.cz
References


References


LVCs Cross-Lingual Comparison Structural Composition of LVCs Morphosyntax of LVCs LVCs in UD Revisited Conclusion
References


