1 Introduction

Case markers express semantic roles, describing the relationship between the arguments they apply to and the action of a verb. Adpositions (prepositions and postpositions) further express a range of semantic relations, including space, time, possession, properties, and comparison.

The use of specific case markers and adpositions for particular semantic roles is idiosyncratic to every language. This poses problems in many natural language processing tasks such as machine translation (Ratnam et al. 2018, Jha 2017, Ramanathan et al. 2009, Rao et al. 1998) and semantic role labelling (Pal and Sharma 2019, Gupta 2019). Models for these tasks rely on human-annotated corpora as training data, such as the one created for the Hindi-Urdu PropBank (Bhatt et al., 2009), and by Kumar et al. (2019).

There is a lack of corpora in South Asian languages for such tasks. Even Hindi, despite being a resource-rich language, is limited in available labelled data (Joshi et al., 2020). This extended abstract presents the in-progress annotation of case markers and adpositions in a Hindi corpus, employing the cross-lingual SNACS scheme (Semantic Network of Adposition and Case Supersenses; Schneider et al., 2018, 2020). The guidelines we are developing also apply to Urdu.

2 Corpus

The corpus was the entirety of the The Little Prince. Annotation was done by one linguistically-trained native speaker of Hindi during June–July 2020, and guidelines were developed simultaneously. Table 1 contains statistics about the corpus, and Table 2 gives proportions for each label and target.

The final version of the corpus will require multiple annotators and adjudication to resolve disagreements.

### Table 1: Statistics about the corpus.

<table>
<thead>
<tr>
<th>Count</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>16,333</td>
</tr>
<tr>
<td>Targets</td>
<td>2,371</td>
</tr>
<tr>
<td>Case markers</td>
<td>1,988</td>
</tr>
<tr>
<td>Adpositions</td>
<td>383</td>
</tr>
<tr>
<td>Supersenses</td>
<td>2,371</td>
</tr>
<tr>
<td>Scene roles</td>
<td>2,371</td>
</tr>
<tr>
<td>Functions</td>
<td>2,371</td>
</tr>
<tr>
<td>Construals</td>
<td>2,371</td>
</tr>
<tr>
<td>Role = Fxn.</td>
<td>1,330</td>
</tr>
<tr>
<td>Role ≠ Fxn.</td>
<td>1,041</td>
</tr>
</tbody>
</table>

### Annotation targets

Following Masica (1993)’s analysis of Indo-Aryan languages, we annotated the Layer II and III function markers in Hindi. These include all of the simple case markers\(^1\) and all of the adpositions.\(^2\) The ubiquitous adjectival suffix \(\bar{v}\ddot{a}l\ddot{a}\) and the comparison terms \(\ddot{ja}i\ddot{s}\ddot{a}\) and \(\ddot{ja}i\ddot{s}\ddot{e}\) were annotated.

The directly-declined Layer I cases of nominative (which is unmarked), oblique, and vocative were not annotated. The final corpus will investigate these further.

3 Applying SNACS to Hindi-Urdu

Several linguistic features of Hindi-Urdu adposition and case semantics posed difficulties in annotating. Some are examined below, and will need to be resolved for a final corpus.

### Functions for case markers

SNACS has adopted a construal system (Hwang et al., 2017) that labels both the semantic role expressed between the governor and the object (\textbf{scene role}) and

\(^{1}\)\textit{ne} (ergative), \textit{ko} (dative-accusative), \textit{se} (instrumental-ablative-comitative), \(\ddot{k}\ddot{a}\ddot{k}\ddot{e}/\ddot{k}\ddot{e}\) (genitive), \textit{men} (locative-IN), \textit{tak} (allative), \textit{par} (locative-ON). Declined forms of the pronouns (including the reflexive \textit{ap\ddot{n}\ddot{a}}) were also included.

\(^{2}\)An open class, given the productivity of the oblique genitive \textit{ke} as a postposition former.
the literal semantics encoded in the choice of adposition (function). For example, a **Recipient** scene role may be framed with an **Agent** function (“I took it”) or a **Theme** function (“He gave it to me”).

Case markers encode less lexical content than adpositions. Table 2 shows the dominance of case markers in every category; given their versatility, delineating their prototypical functional type is difficult. For example, the prototypical way to express a comparative in Hindi-Urdu is with ("like") (see Table 2). Non-nominative/ergative subjects The **Agent** is prototypically expressed with the ergative case marker **ne** or the unmarked nominative. To express modality, Hindi-Urdu, like other Indo-Aryan languages, employs various aspectual light verbs along with differential subject marking (de Hoop and Narasimhan, 2005). One example is the dative subject indicating obligation:

```
(1) a. main- ne likhā
   1SG-ERG write.PRF
   'L-Originator→Agent wrote it.'

b. mujh- ko likhnā parā
   1SG.OBL-DO fall,PRF
   'L-Originator→? had to write it.'
```

In these, the subject’s role is **originator** as it is a producer of writing. In 1b, an expression of obligation, the subject is not only compelled to act by some outer force (fitting a **Theme**) but is also performing the action unaided (**Agent**). SNACS currently cannot resolve the conflict between these two equally valid functions.

Other unconventional subjects are less problematic. South Asian languages near-universally have dative subject **Experiencers** (Verma and Mohanan, 1990). For these, the prototypical **Recipient** subject is fitting. The passive subject also has the unambiguous function of **Agent**.

**Causative constructions** Indo-Aryan languages, through suffixation, derive indirect and direct causative verbs from intransitive verbs. Indirect causatives take an argument in the instrumental case that is an *impelled agent*, grammatically distinguished from a true **Instrument**:

```
(2) us-ne cābhī=se darvāzā kholā
   3SG.ERG key.OBL=INS door.NOM open,PRF
   'She opened the door with a key.'

(3) us-ne mālik=se darvāzā
   3SG.ERG owner.OBL=INS door.NOM
   kholvā,
   open.CAUS,PRF
   'She made the landlord open the door.'
```

Much like an obligated agent, the impelled agent takes part in two events, exhibiting properties of both **Agent** and **Theme**. Furthermore, an impelled agent can control **Instruments** of its own, and there cannot be two participants in the scene with the same semantic role (Begum and Sharma, 2010). For SNACS, Shalev et al. (2019) mentioned similar issues in English.

### 4 Conclusion

We have adapted SNACS to Hindi-Urdu, developing guidelines and annotating a substantial preliminary corpus of *The Little Prince* in Hindi. Issues in annotating case markers, modality, and causatives were raised. Future work will finalize the corpus, resolve these linguistic issues, and examine NLP applications of the data to semantic role labelling and machine translation of adpositions and case markers.

---

Table 2: Breakdown of label counts along various dimensions, divided between case markers (above divider) and adpositions (below divider). Each column is independent and covers the whole dataset.

<table>
<thead>
<tr>
<th>Type</th>
<th>%</th>
<th>Function %</th>
<th>Scene role %</th>
</tr>
</thead>
<tbody>
<tr>
<td>kā (GEN)</td>
<td>28.7</td>
<td>AGENT</td>
<td>EXPERIENCER</td>
</tr>
<tr>
<td>ko (ACC/DAT)</td>
<td>19.1</td>
<td>GESTALT</td>
<td>ORIGINATOR</td>
</tr>
<tr>
<td>ne (ERG)</td>
<td>12.1</td>
<td>THEME</td>
<td>THEME</td>
</tr>
<tr>
<td>se (INS/ABL/COM)</td>
<td>10.7</td>
<td>TOPIC</td>
<td>TOPIC</td>
</tr>
<tr>
<td>mem (LOC-in)</td>
<td>7.6</td>
<td>LOCUS</td>
<td>LOCUS</td>
</tr>
<tr>
<td>par (LOC-on)</td>
<td>4.6</td>
<td>GESTALT</td>
<td>GESTALT</td>
</tr>
<tr>
<td>tak (ALL)</td>
<td>1.0</td>
<td>AGENT</td>
<td>AGENT</td>
</tr>
<tr>
<td>ke lie (“for”)</td>
<td>4.0</td>
<td>COMPRef.</td>
<td>COMPRef.</td>
</tr>
<tr>
<td>jaisa (“like”)</td>
<td>1.3</td>
<td>PURPOSE</td>
<td>PURPOSE</td>
</tr>
<tr>
<td>ke pās (“near”)</td>
<td>1.2</td>
<td>EXPLANATION</td>
<td>EXPLANATION</td>
</tr>
<tr>
<td>kt taraah (“like”)</td>
<td>1.1</td>
<td>MANNER</td>
<td>MANNER</td>
</tr>
<tr>
<td>vālā (adjectival)</td>
<td>1.0</td>
<td>TIME</td>
<td>TIME</td>
</tr>
</tbody>
</table>

---

Some South Asian languages have dative **Possessors**.
References


