SNACS Annotation of Case Markers and Adpositions in Hindi

Aryaman Arora  
*Georgetown University*, aa2190@georgetown.edu

Nitin Venkateswaran  
*Georgetown University*, nv214@georgetown.edu

Nathan Schneider  
*Georgetown University*, nathan.schneider@georgetown.edu

Follow this and additional works at: [https://scholarworks.umass.edu/scil](https://scholarworks.umass.edu/scil)

Part of the Computational Linguistics Commons

Recommended Citation  
Arora, Aryaman; Venkateswaran, Nitin; and Schneider, Nathan (2021) "SNACS Annotation of Case Markers and Adpositions in Hindi," *Proceedings of the Society for Computation in Linguistics*: Vol. 4 , Article 57. DOI: [https://doi.org/10.7275/hfme-g116](https://doi.org/10.7275/hfme-g116)  
Available at: [https://scholarworks.umass.edu/scil/vol4/iss1/57](https://scholarworks.umass.edu/scil/vol4/iss1/57)

This Extended Abstract is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Proceedings of the Society for Computation in Linguistics by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
Abstract

We present in-progress annotation of semantic relations expressed through adpositions and case markers in a Hindi corpus. We used the multilingual SNACS annotation scheme, which has been applied to a variety of typologically diverse languages. Annotation problems in Hindi are examined and used to suggest changes to SNACS. We look towards finalizing the corpus and using it for future work in typology and semantic role-dependent tasks.

1 Introduction

Case markers express semantic roles, describing the relationship between the arguments they apply to and the action of a verb. Adpositions (prepositions, postpositions, and circumpositions) 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. Hindi–Urdu has a case-marking system along with a large postposition inventory. Idiosyncratic bundling of case and adpositional relations poses problems in many natural language processing tasks for Hindi, 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). Many 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). The study of adposition and case semantics in corpora is also useful from a linguistic perspective, in comparing and categorizing the encoding of such relations across languages.

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 reports on 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, since the grammatical base of Hindi and Urdu is largely the same.

2 Corpus

The corpus was the entirety of the The Little Prince. Annotation was done by two highly proficient Hindi speakers (one native), and guidelines were developed simultaneously. Table 1 contains statistics about the corpus, and table 2 gives proportions for each label and target.

Adjudication of annotator disagreements is ongoing and is expected to be completed by February 2021.

Annotation targets Following Masica’s (1993) analysis of Indo-Aryan languages, we annotated the Layer II and III function markers in Hindi.

Table 1: Statistics about the corpus.

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>16,333</td>
<td></td>
</tr>
<tr>
<td>Targets</td>
<td>2,371</td>
<td>55</td>
</tr>
<tr>
<td>Case markers</td>
<td>1,988</td>
<td>6</td>
</tr>
<tr>
<td>Adpositions</td>
<td>383</td>
<td>51</td>
</tr>
<tr>
<td>Supersenses</td>
<td>2,371</td>
<td>50</td>
</tr>
<tr>
<td>Scene roles</td>
<td>2,371</td>
<td>48</td>
</tr>
<tr>
<td>Functions</td>
<td>2,371</td>
<td>41</td>
</tr>
<tr>
<td>Construals</td>
<td>2,371</td>
<td>143</td>
</tr>
<tr>
<td>Role = Fxn.</td>
<td>1,330</td>
<td>38</td>
</tr>
<tr>
<td>Role ≠ Fxn.</td>
<td>1,041</td>
<td>105</td>
</tr>
</tbody>
</table>
These include all of the simple case markers\(^2\) and all of the adpositions.\(^3\)

We also decided to annotate the suffix vālā when used in an adjectival sense (e.g. chōṭā-vālā kambrā ‘the room that is small’), the comparison terms jaisā and jaise, the extent and similarity particle sā (chōṭā-sā kambrā ‘small-ish room’), and the emphatic particles bhī, hī, to (Koul, 2008, 137–156). All of these modify the preceding token and mediate a semantic relation between their object and the object’s governor, just as conventionally-designated postpositions do.

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

### 3 SNACS

The Semantic Network of Adposition and Case Supersenses (SNACS; Schneider et al., 2018, 2020) is a multilingual annotation scheme with 50 supersenses that characterize the use of adpositions and case markers at a coarse level of granularity. This scheme is akin to linguistic models of argument structure such as semantic roles and theta roles (including traditional categories such as AGENT and THEME), but expanded to include roles for adpositional relations, such as WHOLE for whole–part, SOCIAL.REL for interpersonal relations, etc.

A useful feature of SNACS is the construal system (Hwang et al., 2017), which allows an annotator to give one label for the morphosyntactic role or inherent lexical meaning (function) and another label for the predicate-licensed semantic relation (scene role) of a token. This is expressed as SCENE ROLE→FUNCTION, and is useful for disambiguating the use of different encodings of the same semantic relation such as “she is looking at me” (STIMULUS→DIRECTION) and “he is listening to me” (STIMULUS→GOAL). When the scene role and function are identical, a single label is given.

SNACS, thus far, has been used to annotate the English STREUSLE corpus (Schneider and Smith, 2015), The Little Prince in English and translations of it into Korean (Hwang et al., 2020), Mandarin (Peng et al., 2020; Zhu et al., 2018) and German, with ongoing annotation efforts in Finnish, Latin, and Gujarati and past work on French and Hebrew. There has also been annotation of L2 English (Kranzlein et al., 2020). This effort is accompanied by the release of language-specific guidelines (based on Schneider et al., 2020) that aid in annotator training.

### 4 Applying SNACS to Hindi–Urdu

Several linguistic features of Hindi–Urdu adposition and case semantics posed difficulties in annotating. Some are examined below. The annotation process itself relied on grammatical analyses of Hindi such as Koul (2008), dictionaries (McGregor, 1993; Dasa, 1965–1975), and native speaker judgements.

#### Functions for case markers

Case markers encode little lexical content relative to adpositions. Table 2 shows the dominance of case markers in every category; given their versatility, delineating their prototypical functions is difficult. For exam-

---

\(^2\)ne (ergative), ko (dative-accusative), se (instrumental-ablative-comitative), kālīkātī (genitive), mem (locative-IN), tak (allative), par (locative-ON). Declined forms of the pronouns (including the reflexive apnā) were also included.

\(^3\)An open class, given the productivity of the oblique genitive ke as a postposition former.

---

Table 2: Breakdown of label counts along various dimensions, divided between case markers and adpositions. Each of the 8 tables is independent.
ple, a comparative in Hindi–Urdu is expressed with
the ablative case marker se—should the function
be SOURCE (as expected for the ablative case) or
the narrower COMPARISON in this sense? This
is an unresolved question; in labelling, we chose
narrower functions when their use seemed to be
a relation that is not completely supplied by the
predicate.

In other cases, with highly polysemous markers
such as se, it is difficult to pick a single function
corresponding to an obvious grammatical case. For
example, the verb pūchnā ‘to ask’ takes an argu-
ment, marked with se, indicating the person being
asked. This instance of se could be construed as
the ablative case (reflecting the return of a response
from the person asked) or the comitative case (indicat-
ing a co-participant in communication, exactly
as for verbs such as kahnā ‘to say’).

To resolve this issue we looked to typological evi-
dence, in keeping with SNACS’s multilingual aims:
the closely-related language Punjabi, which has
separate ablative (tom) and comitative (nāl) marks-
ers, uses the ablative in this construction, so we
labelled the function SOURCE.

Non-nominative/ergative subjects  The AGENT
is prototypically expressed with the ergative case
marker ne or the unmarked nominative. To ex-
press modality, Hindi–Urdu, like other Indo-Aryan
languages, employs various aspecl actual light verbs
along with differential subject marking (de Hoop
and Narasimhan, 2005). One example is the dative
subject indicating obligation:

(1) us-se apnā savāl pūcho.
    3SG.OBL.-? self.GEN question ask.IMPF
‘Ask them:RECIPIENT~? your question.’

Much like an obligated agent, the impelled agent
takes part in two events, exhibiting properties of
both AGENT and THEME. Furthermore, an im-
elled 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.

This construction was rare in our corpus, but we
find the best solution for this is a new label for
animate and ambiguously volitional counterparts
to INSTRUMENT in the SNACS hierarchy, much
like the distinction between inanimate CAUSER and
animate AGENT.

Emphatic particles  Following work on SNACS
for Korean, which created a new label FOCUS
for “postpositions that indicate the focus of a sen-
tence (FOC), contributing information such as con-
trastiveness, likelihood, or value judgements”
(Hwang et al., 2020), we found that the Hindi em-
phatic particles hī ‘only’, bhī ‘also, too’, to (con-
trastive), and some uses of tak ‘even’ function as
focus postpositions and thus merited annotation.

2Some South Asian languages have dative POSSESSORS.
Table 3: Raw interannotator agreement statistics for all targets (top third), and the top three (middle third) and bottom three (bottom third) targets by scene role and function agreement.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Scene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>2,371</td>
<td>75.7%</td>
<td>82.4%</td>
</tr>
<tr>
<td>Top 5</td>
<td>1,856</td>
<td>76.9%</td>
<td>84.4%</td>
</tr>
<tr>
<td>Other</td>
<td>515</td>
<td>71.6%</td>
<td>75.6%</td>
</tr>
<tr>
<td>ke bāre mem.</td>
<td>23</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>ke alāvā</td>
<td>7</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>ke lie</td>
<td>95</td>
<td>89.5%</td>
<td>96.8%</td>
</tr>
<tr>
<td>ke pratī</td>
<td>8</td>
<td>0.0%</td>
<td>87.5%</td>
</tr>
<tr>
<td>tak</td>
<td>23</td>
<td>73.9%</td>
<td>56.5%</td>
</tr>
<tr>
<td>ke pās</td>
<td>31</td>
<td>93.5%</td>
<td>51.6%</td>
</tr>
</tbody>
</table>

Table 4: Cohen’s Kappa and Fleiss’ Kappa statistics for measuring inter-rater reliability.

<table>
<thead>
<tr>
<th></th>
<th>Scene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohen’s κ</td>
<td>0.7469</td>
<td>0.8104</td>
</tr>
<tr>
<td>Fleiss’ κ</td>
<td>0.7504</td>
<td>0.8164</td>
</tr>
</tbody>
</table>

5 Interannotator agreement

Table 3 shows interannotator agreement rates over targets for all chapters of The Little Prince annotated by both annotators. The Top 5 category includes the top five case markers by their label counts, as seen in table 2. The table also shows interannotator agreement rates for the top and bottom three target types by scene and function agreement. Some targets have unambiguous lexical and governor-licensed meanings, such as ke bāre mem (marked TOPIC for both) and ke alāvā (marked PARTITION for both). The case of ke pratī is somewhat unusual in having zero scene agreement with high function agreement (with the function largely agreed as DIRECTION), suggesting a versatility in the interpretation of the governor-licensed relationship. The case of ke pās has high scene agreement (given that the term unambiguously indicates the possessor-possession relationship between governor and governee) and low function agreement (due to the annotators’ disagreement over the possessive or locative significance of the term).

Table 4 shows Cohen’s κ (Cohen, 1960) and Fleiss’ κ (Fleiss, 1971) inter-annotator agreement statistics for the scene and function roles. The probabilities of agreement by chance using Cohen’s κ metric are 4.0% for the scene label and 7.1% for the function label. The probabilities of agreement by chance using Fleiss’ κ metric are 5.1% for the scene label and 7.2% for the function label. These low probabilities suggest the presence of well-defined patterns of lexical and governor-licensed meanings of case markers and adpositions.

6 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, such as automatic prediction of SNACS labels, alignment and cross-lingual comparison, and the release of guidelines for Hindi–Urdu.

Acknowledgements

We thank Jena Hwang, Vivek Srikumar, and Jakob Prange for helpful discussions about annotation issues and drafting guidelines, Miriam Butt for suggesting useful linguistic literature, and Archna Bhatia for initial work on Hindi before this annotation project began. We also thank the three anonymous SCiL reviewers and three more anonymous SIGTYP reviewers whose input was invaluable in writing this extended abstract.

References


