SNACS Annotation of Case Markers and Adpositions in Hindi

Aryaman Arora **Nitin Venkateswaran** **Nathan Schneider**
Georgetown University
{aa2190,nv214,nathan.schneider}@georgetown.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.

	Count	Types
Tokens	16,333	
Targets	2,371	55
Case markers	1,988	6
Adpositions	383	51
Supersenses	2,371	50
Scene roles	2,371	48
Functions	2,371	41
Construals	2,371	143
Role = Fxn.	1,330	38
Role ≠ Fxn.	1,041	105

Table 1: Statistics about the corpus.

2 Corpus

The corpus was the entirety of the *The Little Prince*.[1] 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.

[1]The corpus is available at https://github.com/aryamanarora/carmls-hi.

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 454-458.
Held on-line February 14-19, 2021

	Type	%	*Scene role*	%	*Function*	%	*Scene role⇝Function*	%
Case Markers	*kā* (GEN)	28.7	EXPERIENCER	11.1	AGENT	13.4	THEME⇝THEME	6.7
	ko (ACC/DAT)	19.1	ORIGINATOR	8.3	GESTALT	11.9	EXPERIENCER⇝RECIPIENT	6.4
	ne (ERG)	12.1	THEME	7.3	THEME	11.3	ORIGINATOR⇝AGENT	5.9
	se (INS/ABL/COM)	10.7	TOPIC	6.5	RECIPIENT	9.0	LOCUS⇝LOCUS	5.5
	meṃ (LOC-in)	7.6	LOCUS	6.0	LOCUS	7.6	GESTALT⇝GESTALT	5.1
	par (LOC-on)	4.6	GESTALT	5.4	SOURCE	5.1	LOCUS⇝LOCUS	4.7
	tak (ALL)	1.0	AGENT	5.2	TOPIC	4.6	AGENT⇝AGENT	4.1
	Type	%	*Scene role*	%	*Function*	%	*Scene role⇝Function*	%
Adpositions	*ke lie* ("for")	4.0	COMPREF.	2.3	COMPREF.	3.0	COMPREF.⇝COMPREF.	2.2
	jaise ("like")	1.3	PURPOSE	1.3	BENEFICIARY	1.6	PURPOSE⇝PURPOSE	1.3
	ke pās ("near")	1.2	EXPLANATION	1.3	LOCUS	1.4	EXPL.⇝EXPL.	1.3
	kī tarah ("like")	1.1	MANNER	1.3	PURPOSE	1.4	EXPERIENCER⇝BENEF.	1.1
	vālā (adjectival)	1.0	TIME	1.1	EXPLANATION	1.3	TOPIC⇝TOPIC	1.0

Table 2: Breakdown of label counts along various dimensions, divided between case markers and adpositions. **Each of the 8 tables is independent.**

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. *choṭā-vālā kamrā* 'the room that is small'), the comparison terms *jaisā* and *jaise*, the extent and similarity particle *sā* (*choṭā-sā kamrā* '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, SOCIALREL 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 SCENEROLE⇝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), *kā*/*ke*/*kī* (genitive), *meṃ* (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.

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 COMPARISONREF 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 argument, 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 (indicating a co-participant in communication, exactly as for verbs such as *kahnā* 'to say').

(1) us-**se** apnā savāl pūcho.
3SG.OBL-? self.GEN question ask.IMP
'Ask them:RECIPIENT⇝? your question.'

To resolve this issue we looked to typological evidence, in keeping with SNACS's multilingual aims: the closely-related language Punjabi, which has separate ablative (*toṃ*) and comitative (*nāl*) markers, 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 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:

(2) a. maiṃ-**ne** likhā
1SG-ERG write.PRF
'I:ORIGINATOR⇝AGENT wrote it.'

b. mujh-**ko** likhnā paṛā
1SG.OBL-DAT do.INF fall.PRF
'I:ORIGINATOR⇝? had to write it.'

In these, the subject's scene role is ORIGINATOR as it is a producer of writing. In (2b), 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; we currently label (2b) as ORIGINATOR⇝RECIPIENT in keeping with the morphosyntax of the dative subject. The issue is a broader problem of dealing with force dynamics in semantic role labelling, and may require new labels.

Other unconventional subjects are less problematic. South Asian languages near-universally have dative subject EXPERIENCERs (Verma and Mohanan, 1990).[4] For these, the prototypical RECIPIENT subject is fitting. The passive subject also has the unambiguous function of AGENT, just as the English passive **by**.

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:

(3) us-ne cābhī=**se** darvāzā kholā
3SG.ERG key.OBL=INS door.NOM open.PRF
'She opened the door [with a key]:INSTRUMENT.'

(4) us-ne mālik=**se** darvāzā khulvāyā
3SG.ERG owner.OBL=INS door.NOM 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.

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 sentence (FOC), contributing information such as contrastiveness, likelihood, or value judgements" (Hwang et al., 2020), we found that the Hindi emphatic particles *hī* 'only', *bhī* 'also, too', *to* (contrastive), and some uses of *tak* 'even' function as focus postpositions and thus merited annotation.

[4] Some South Asian languages have dative POSSESSORs.

Category	Count	Scene	Function
All	2,371	75.7%	82.4%
Top 5	1,856	76.9%	84.4%
Other	515	71.6%	75.6%
ke bāre meṃ	23	100.0%	100.0%
ke alāvā	7	100.0%	100.0%
ke lie	95	89.5%	96.8%
ke prati	8	0.0%	87.5%
tak	23	73.9%	56.5%
ke pās	31	93.5%	51.6%

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.

	Scene	Function
Cohen's κ	0.7469	0.8104
Fleiss' κ	0.7504	0.8164

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

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 meṃ* (marked TOPIC for both) and *ke alāvā* (marked PARTPORTION for both). The case of *ke prati* 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

Rafiya Begum and Dipti Misra Sharma. 2010. A preliminary work on Hindi causatives. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 120–128, Beijing, China.

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Sharma, and Fei Xia. 2009. A multi-representational and multi-layered treebank for Hindi/Urdu. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 186–189, Suntec, Singapore.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37 – 46.

Syamasundara Dasa. 1965–1975. *Hindī śabdasāgara*. Nagari Pracarini Sabha.

Helen de Hoop and Bhuvana Narasimhan. 2005. Differential case-marking in Hindi. In Mengistu Amberber and Helen De Hoop, editors, *Competition and Variation in Natural Languages*, Perspectives on Cognitive Science, pages 321 – 345. Elsevier, Oxford.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378 – 382.

Aishwary Gupta. 2019. *Semantic Role Labeling for Indian languages*. Ph.D. thesis, International Institute of Information Technology Hyderabad.

Jena D. Hwang, Archna Bhatia, Na-Rae Han, Tim O'Gorman, Vivek Srikumar, and Nathan Schneider. 2017. Double trouble: The problem of construal in semantic annotation of adpositions. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 178–188, Vancouver, Canada.

Jena D. Hwang, Hanwool Choe, Na-Rae Han, and Nathan Schneider. 2020. K-SNACS: Annotating Korean adposition semantics. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 53–66, Barcelona Spain (online). Association for Computational Linguistics.

Sanjay Kumar Jha. 2017. Translation of English Prepositions into Hindi Postpositions. *International Journal of Innovations in TESOL and Applied Linguistics*, 3(4).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online.

Omkar N. Koul. 2008. *Modern Hindi Grammar*. Dunwoody Press.

Michael Kranzlein, Emma Manning, Siyao Peng, Shira Wein, Aryaman Arora, and Nathan Schneider. 2020. PASTRIE: A corpus of prepositions annotated with supersense tags in Reddit international English. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 105–116, Barcelona, Spain. Association for Computational Linguistics.

Ritesh Kumar, Bornini Lahiri, and Atul Kr. Ojha. 2019. Cross-linguistic semantic tagset for case relationships. In *Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP*.

Colin P. Masica. 1993. *The Indo-Aryan Languages*. Cambridge University Press.

R. S. McGregor. 1993. *The Oxford Hindi-English dictionary*. Oxford University Press.

Riya Pal and Dipti Sharma. 2019. A dataset for semantic role labelling of Hindi-English code-mixed tweets. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 178–188, Florence, Italy.

Siyao Peng, Yang Liu, Yilun Zhu, Austin Blodgett, Yushi Zhao, and Nathan Schneider. 2020. A corpus of adpositional supersenses for Mandarin Chinese. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5986–5994, Marseille, France. European Language Resources Association.

Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. 2009. Case markers and morphology: Addressing the crux of the fluency problem in English-Hindi SMT. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 800–808, Suntec, Singapore.

D. Rao, P. Bhattacharya, and Radhika Mamidi. 1998. Natural language generation for English to Hindi human-aided machine translation. *Proceedings of the International Conference on Knowledge Based Computer Systems*.

D. Jyothi Ratnam, M. Anand Kumar, B. Premjith, K. P. Soman, and S. Rajendran. 2018. Sense disambiguation of English simple prepositions in the context of English–Hindi machine translation system. In S. Margret Anouncia and Uffe Kock Wiil, editors, *Knowledge Computing and Its Applications: Knowledge Manipulation and Processing Techniques*, volume 1, pages 245–268. Springer, Singapore.

Nathan Schneider, Jena D. Hwang, Archna Bhatia, Vivek Srikumar, Na-Rae Han, Tim O'Gorman, Sarah R. Moeller, Omri Abend, Adi Shalev, Austin Blodgett, and Jakob Prange. 2020. Adposition and Case Supersenses v2.5: Guidelines for English. *arXiv:1704.02134 [cs]*.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proc. of ACL*, pages 185–196, Melbourne, Australia.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547, Denver, Colorado. Association for Computational Linguistics.

Adi Shalev, Jena D. Hwang, Nathan Schneider, Vivek Srikumar, Omri Abend, and Ari Rappoport. 2019. Preparing SNACS for subjects and objects. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 141–147, Florence, Italy.

Mahendra K. Verma and Karuvannur Puthanveettil Mohanan. 1990. *Experiencer subjects in South Asian languages*. Center for the Study of Language (CSLI).

Yilun Zhu, Yang Liu, Siyao Peng, Austin Blodgett, Yushi Zhao, and Nathan Schneider. 2018. Adpositional supersenses for mandarin chinese. *ArXiv*, abs/1812.02317.