SNACS Annotation of Case Markers and Adpositions in Hindi

Aryaman Arora
Georgetown University
aa2190@georgetown.edu

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

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.

	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.

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 *vālā* and the comparison terms *jaisā* and *jaise* 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 (**scene role**) and

[1] *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.

[2] An open class, given the productivity of the oblique genitive *ke* as a postposition former.

	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
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 (above divider) and adpositions (below divider). **Each column is independent and covers the whole dataset.**

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 prototypical functions is difficult. For example, the prototypical way to express a comparative in Hindi-Urdu is with the ablative case—should the function be SOURCE or COMPARISONREF in this sense? This is an unresolved question; in labelling, we chose narrower functions when possible.

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. 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 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).[3] 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 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.

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.

[3] Some South Asian languages have dative POSSESSORs.

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.

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.

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.

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.

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.

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.

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.

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).