MASALA: Modelling and Analysing the Semantics of Adpositions in Linguistic Annotation of Hindi

Aryaman Arora, Nitin Venkateswaran, Nathan Schneider
Georgetown University
Washington, D.C., USA
{aa2190, nv214, nathan.schneider}@georgetown.edu

Abstract

We present a completed, publicly available corpus of annotated semantic relations of adpositions and case markers in Hindi. We used the multilingual SNACS annotation scheme, which has been applied to a variety of typologically diverse languages. Building on past work examining linguistic problems in SNACS annotation, we use language models to attempt automatic labelling of SNACS supersenses in Hindi and achieve results competitive with past work on English. We look towards upstream applications in semantic role labelling and extension to related languages such as Gujarati.

Keywords: SNACS, semantic parsing, Hindi

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.

Languages have different strategies for encoding these kinds of semantic relations. Hindi–Urdu[1] uses a case-marking system along with a large postposition inventory (Kachru, 2006; Koul, 2008). 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 for training data, such as the one created for the Hindi–Urdu PropBank (Bhatt et al., 2009), and in Kumar et al. (2019). The study of adposition and case semantics in corpora is also useful from a comparative/typological linguistic perspective, in comparing and categorizing the encoding of such relations across languages.

To that end, we release a completed Hindi corpus annotated for adposition and case semantic labels using the SNACS formalism (Schneider et al., 2018a, 2020). We approach the problem of automatic tagging of these labels using a variety of language models and explore what these models learn. Drawing on parallel SNACS corpora in English, German, Mandarin, and Korean, we compare strategies for encoding semantic roles across languages.

2. Background

Hindi is a language of India, of the Indo-Aryan branch of the Indo-European family, and one of the best-resourced South Asian languages for research in natural language processing and computational linguistics (Joshi et al., 2020). Hindi has a small number of core case markers as well as a large class of adpositions for signalling semantic relations. We will discuss the linguistic features of case and adposition in Hindi below, related work from linguistics in this area, and introduce the SNACS schema.

2.1. Case and adposition in Hindi

Hindi is generally described as having three layers of case/adposition: the three basic morphological cases (example 1a), a small class of case markers/clitics that indicate core arguments to verbs (example 1b), and a larger class of postpositions governed by the genitive *kā* or ablative *se* (example 1c) (Kachru, 2006).

(1) a. bacc**e** ‘children’, bacc**oṁ** ‘children.OBL’, bacc**o** ‘children.VOC’

b. us**ne** ‘she.ERG’, us**ko** ‘she.ACC/DAT’

c. us**ke_liye** ‘for her’, us**ke_nazdīk** ‘near her’, us**ke_*under*** ‘under her’ [code-switching]

Masica (1993) grouped these three “layers” on the basis of historical development. Diachronically, morphological cases are the remnants of the Indo-European case system (via Sanskrit) that largely encode syntactic information, the case markers are Middle Indo-Aryan developments from spatial adverbs (e.g. Sanskrit *upari* ‘above’ > Hindi *par* ‘LOC-on’) that encode fundamental semantic roles on complements, and postpositions are more recent developments that even include borrowings from Persian, Arabic, and English and which indicate more concrete, e.g. spatial, relations between nominals.

The case markers most commonly mark relations between verbs and their arguments and adjuncts, followed by relations between nominals. Case markers in Hindi

[1] Hindi and Urdu are two registers written in two different scripts of a single language (usually called ‘Hindi–Urdu’ or ‘Hindustani’) with a largely identical grammar. While our corpus is in Hindi in the Devanagari script, the linguistic portions of our work (e.g. annotation guidelines) are applicable to Urdu as well.

are highly multi-functional even when using coarse descriptors from linguistic typology; e.g. *se* is described as indicating the ablative, instrumental, comitative, or comparative cases depending on context, respectively exemplified in (2).

(2) a. yahāṁ **se** jāō ‘Go away **from** here.’
b. cammac **se** k^{h}ānā ‘eating **with** a spoon’
c. us**se** milūṁgā ‘I will meet **with** him.’
d. das **se** kam ‘less **than** ten’

That is not to say that *se* is three different case markers; the semantic role of a *se*-marked argument is just licensed by the predicate or other governor of the argument. Understanding how and in what context such markers indicate what semantic relations is an interesting problem. Thus far, there is no semantically-annotated corpus of case and postposition semantics in Hindi, which motivated our annotation of this corpus.

2.2. Related work

There is a great deal of work on case and adpositions in Hindi. In syntax, some research topics are syntactic differences between morphological case, case markers, and adpositions (Spencer, 2005), the issue of differential case marking in the ergative and dative–accusative (Bhatt and Anagnostopoulou, 1996; de Hoop and Narasimhan, 2005; de Hoop and Narasimhan, 2009; Montrul et al., 2015; Montaut, 2018), word order (Mohanan, 1994), and agreement (Montrul et al., 2012).

On the other hand, there has been less research on the semantics of case and adpositions in Hindi. The mapping of case-marked arguments to lexical-semantic roles has been done in various computational projects (Begum et al., 2008; Vaidya et al., 2011). Paul et al. (2010) is an investigation of paraphrasing nominal compound relations with case in Hindi and English.

2.3. SNACS

The Semantic Network of Adposition and Case Supersenses (SNACS; Schneider et al., 2018a, 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 if they differ. Examples of SNACS annotation for Hindi are given below.

	Count	%	Types
CHAPTERS	27		
SENTENCES	1,580		
TOKENS	16,882		
TARGETS	2,970		70
Case	2,142	72.1%	7
Emphatic	382	12.9%	3
Adpositions	446	15.0%	60
CONSTRUALS	2,970		136
Role = Fxn.	1,886	63.4%	38
Role ≠ Fxn.	1,084	36.6%	98

Table 1: Cumulative statistics of the Hindi corpus.

(3) vah g^{h}ar **ke_pās**$_{\text{LOCUS}}$ hai
3SG home near COP.IND.3SG
‘He is near the house.’

(4) maiṁ us **ko**$_{\text{THEME}}$ k^{h}ā-tā hūṁ
1SG 3SG ACC eat-IPFV.M.SG COP.IND.1SG
‘I eat that.’

(5) maiṁ **ne**$_{\text{EXPERIENCER⇝AGENT}}$ nadī
1SG ERG river
ke_pār$_{\text{LOCUS⇝PATH}}$ ek baccā dekh-ā
across one child.NOM see-PFV.M.SG
‘I saw a child across the river.’

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), and German (Prange and Schneider, 2021). There has also been annotation of L2 English (Kranzlein et al., 2020). This effort has been accompanied by the release of guidelines for annotator training, including for English (Schneider et al., 2020) and Hindi–Urdu (Arora et al., 2021a). Some earlier works also discussed linguistic issues in Hindi annotation (Arora et al., 2021b).

There is also an online interface for exploring SNACS corpora and interactive annotation guidelines: `http://www.xposition.org/` (Gessler et al., 2022).

3. Corpus and annotation

The corpus was the entirety of *Nanhā Rājkumār*, the Hindi translation of the *The Little Prince* by Antoine de Saint-Exupéry.[2] We used the SNACS annotation scheme, of which a brief overview is given in §2.3. Annotation was done by two Hindi speakers: A (the first author, who is a native speaker) and B (the second author, who is highly proficient) during June 2020–January 2021, and annotation guidelines were developed simultaneously (Arora et al., 2021a). Table 1 contains statistics about the final corpus, which was released in CoNLL-U-Lex format with Universal Dependencies annotations generated with Stanza (Qi et al., 2020).

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

There were two phases of annotation. In the first, A annotated the whole corpus (including all case markers and adpositions) and developed basic guidelines. In the second, B annotated chapter-by-chapter and A and B adjudicated disagreements concurrently. B also annotated focus markers, which were not included as targets in the first phase. A final pass was then conducted over the whole corpus to reconcile any remaining annotation disagreements.

3.1. Annotation targets

Following Masica's (1993) analysis of Indo-Aryan languages, we annotated the Layer II and III function markers in Hindi. These include all of the simple case markers[3] and all of the adpositions.

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 due to the much greater annotation load that would involve and how much greater the breadth of the annotations would be relative to other SNACS-annotated languages. This means verbal arguments without case clitics were not annotated. However, future work (especially with application to semantic role labelling) would benefit from such annotations, and similar work has been done on SNACS annotation of non-adpositionally-marked subjects and objects in English (Shalev et al., 2019).

3.2. Linguistic issues

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 example, 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').

(6) 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 (*toṁ*) 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:

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

b. mujh-**ko** likhnā paṛā
1SG.OBL-DAT write.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 example 7b, 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 example 7b 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] *ne* (ergative), *ko* (dative-accusative), *se* (instrumental-ablative-comitative), *kā/ke/kī* (genitive), *meṁ* (locative-IN), *tak* (allative), *par* (locative-ON). Declined forms of the pronouns (including the reflexive *apnā*) were also included.

[4] Some South Asian languages also have dative POSSESSORs.

	Type	%	*Scene role*	%	*Function*	%	*Scene role⇝Function*	%
Case Markers	*kā* (GEN)	24.3	Experiencer	8.8	Agent	11.0	Theme⇝Theme	5.7
	ko (ACC/DAT)	15.8	Originator	6.8	Theme	10.2	Experiencer⇝Recipient	5.2
	ne (ERG)	10.2	Theme	6.5	Gestalt	9.9	Originator⇝Agent	4.9
	se (INS/ABL/COM)	9.5	Locus	5.2	Recipient	7.7	Gestalt⇝Gestalt	4.5
	meṁ (LOC-in)	6.3	Topic	5.1	Locus	6.5	Locus⇝Locus	4.0
	par (LOC-on)	5.2	Gestalt	4.7	Source	4.0	Agent⇝Agent	3.4
	tak (ALL)	0.8	Agent	4.7	Topic	3.1	Topic⇝Topic	3.0
Focus	*to* (contrastive)	6.2	Focus	9.6	Focus	9.6	Focus⇝Focus	9.6
	hī (“even”)	3.6	`d	2.9	`d	2.9	`d⇝`d	2.9
	bhī (“also”)	3.0	NONSNACS	0.4	NONSNACS	0.4	NONSNACS⇝NONSNACS	0.4
Adpositions	*ke lie* (“for”)	3.3	CompRef.	2.1	CompRef.	2.8	CompRef.⇝CompRef.	2.1
	ke pās (“near”)	1.0	Purpose	1.1	Locus	1.7	Purpose⇝Purpose	1.1
	sā (“-ish”)	1.0	Explanation	1.1	Beneficiary	1.3	Expl.⇝Expl.	1.1
	kī tarah (“like”)	1.0	Time	1.0	Purpose	1.1	Extent⇝Extent	0.9
	jaise (“like”)	1.0	Extent	0.9	Explanation	1.1	Experiencer⇝Benef.	0.9

Table 2: Breakdown of label counts along various dimensions, divided between case markers and adpositions. **Each of the 8 tables is independent.** (E.g., the topmost ‘Scene role’ table shows that 8.8% of annotated targets in the corpus are case markers with the scene role Experiencer.)

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

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

3.3. Corpus analysis

Annotator agreement 2,368 targets (79.7% of the total) were annotated independently by both annotators. In the first round of annotation by annotator A, the focus markers and a small number of case markers were not annotated.

Target	*n*	**Scene**	**Fxn**	**Cons**
ke bāre meṁ (“about”)	23	1.00	1.00	1.00
ke lie (“for”)	95	0.88	0.96	0.87
ne (ERG)	288	0.89	0.98	0.87
kī tarah (“like”)	29	0.83	0.97	0.83
ko (ACC/DAT)	446	0.83	0.95	0.81
par (LOC-on)	107	0.83	0.86	0.79
meṁ (LOC-in)	180	0.80	0.86	0.77
se (INS/ABL/COM)	253	0.79	0.81	0.68
kā (GEN)	682	0.72	0.79	0.66
jaise (“like”)	28	0.57	0.86	0.54
ke pās (“near”)	30	0.97	0.53	0.53
vālā (adjectival)	22	0.36	0.41	0.36
tak (ALL)	23	0.65	0.43	0.35

Table 3: Raw agreement on targets with at least 20 doubly-annotated instances in the corpus, sorted by agreement on the construal. Case markers are in **bold**.

Cohen’s κ between both annotators for double-annotated targets was 0.78 on scene roles, 0.85 on functions, and 0.73 on construals (role⇝function), all of which are very high even compared to previous work on SNACS. It is not surprising that functions, which are inherent to the target type and less dependent on semantics, are easier to annotate than scene roles.

Table 3 shows raw agreements on high-frequency targets, sorted by agreement on the construal label. Among the case markers, *ne* (ERG) and *ko* (ACC/DAT) are the easiest to annotate, which is unsurprising given that their usage is very consistent syntactically (subjects and objects/indirect objects, respectively). The

Target	$\widehat{H}$	n
se (INS/ABL/COM)	3.90	281
kā (GEN)	3.88	723
meṁ (LOC-in)	3.17	187
par (LOC-on)	2.78	155
ko (ACC/DAT)	2.75	470
ke lie ("for")	2.48	97
ke pās ("near")	2.00	31
jaise ("like")	1.85	29
vālā (adjectival)	1.83	28
tak (ALL)	1.79	24
ne (ERG)	1.64	302
to (contrastive)	1.27	185
kī tarah ("like")	0.74	29
sā ("-ish")	0.47	31
ke bāre meṁ ("about")	0.00	23
bhī ("also")	0.00	90
hī ("even")	0.00	107

Table 4: Estimated entropy of targets with at least 20 instances in the corpus. Case markers are in **bold**.

low agreement on *tak* (ALL, "until, up to") was due to uncertainty over whether it indicates the endpoint of movement (GOAL) or the length of the distance covered to the endpoint (EXTENT); after adjudication, we standardised on the latter. The adposition *ke pās* "near" had a similar problem, where we disagreed on whether POSSESSOR was an inherent syntactic function of it or a semantic extension of its spatial use.

Marker and tag distributions Counts of targets and labels are presented in table 2, which shows that case markers generally indicate core arguments of verbs (e.g. AGENT as in subjects of verbs) and basic spatial relations (LOCUS, SOURCE), focus markers have discourse uses, and adpositions indicate non-core adjuncts (e.g. PURPOSE 'in order to').

Since Hindi has case markers, annotated targets were dominated by a few types with very large semantic breadth. We can operationalise a measure of **semantic range** using the entropy of the distribution of scene role labels, which are a coarse representation of semantics, for each case marker. Given a distribution x (the scene roles) with classes K, Shannon entropy (in bits) is defined as:

$$H(x) = -\sum_{k=1}^{K} p(x_k)\log_2 p(x_k) \tag{1}$$

We further adjust for the sample size and distribution using the entropy estimator due to Chao and Shen (2003), which is suited for linguistic distributions (Arora et al., 2022). In table 4, we report entropy of scene role for adpositions and case markers with at least 20 occurrences in the corpus. Case markers, as expected, occupy the top 5 places. However, *tak* ("until, up to") and ergative-case *ne* are much less semantically diverse than the other case markers. Some of the more frequent adpositions are also very semantically diverse, but most are not and form a long tail.

Language	Model	Scene	Fxn
English	Schneider et al. (2018b)	58.2	66.7
	Liu et al. (2021)	**71.9**	**81.0**
Hindi	Baseline	40.1	56.2
	IndicBERT	41.1	59.0
	mBERT	52.0	68.7
	MuRIL	55.4	70.6
	XLM-R	58.7	74.3
	IT distilBERT	69.1	78.7
	IT BERT	**71.4**	**81.8**

Table 5: F1 scores on Hindi test set (n = 158 sentences), only evaluated on gold and predicted `B` tags, compared with past scores on English SNACS tagging on the STREUSLE corpus. Baseline scores are based on picking the most common tag for a given target. Language models in italics are monolingual.

4. Automatic tagging

Given the recent abundance of language models (both multi- and monolingual) for Hindi, we were interested in how well SNACS labels could be automatically tagged. To that end, we trained a neural sequence tagger on the task of adposition and case marker segmentation and tagging of scene role and function. This is a subinstance of the **lexical semantic recognition** (LSR) task first proposed in Liu et al. (2021), who approached it with models similar to those used for named entity recognition (NER). Our tagger feeds the output of a contextual language model through a biLSTM then to a CRF which emits the final tagging. We loaded language models through HuggingFace (Wolf et al., 2020) and implemented our models with PyTorch (Paszke et al., 2019) and AllenNLP (Gardner et al., 2018).

4.1. Data preparation

In preparation for training a classifier, we converted the SNACS labels to the `BIO` scheme for sequence tagging. We only marked the label to be predicted on the `B`-tag of each sequence. In the case of the `B`-tagged-word being segmented into subwords by the language model being used, we labelled all non-initial subwords as `I`.

For example, using multilingual BERT the phrase *uske pīche* 'behind them (sg.)' is tokenised into subwords and tagged for scene role labels as:

_uske	*_pī*	*cha*	*e*
`B`-Locus	`I`	`I`	`I`

The sentences in the dataset are randomly split 80/10/10 between train/dev/test. Training occurs on the train set with period checks against the development set to measure convergence. Scores are reported on the test set.

4.2. Model

The language models we tested are IndicBERT (Kakwani et al., 2020), the original multilingual BERT (Devlin et al., 2019), MuRIL (Khanuja et al., 2021), XLM-RoBERTa (Conneau et al., 2020), and some of the models from the Indic-Transformers library (Jain et al., 2020).

The outputs from the language models are inputted to a 2-layer biLSTM with dropout of 0.3. Its output goes to a CRF, and the highest probability tags are outputted through Viterbi decoding. The number of epochs trained $\{30, 60\}$, the learning rate $\{0.0001, 0.0002, 0.0005, 0.001\}$, and LSTM layer size $\{64, 128, 256, 512\}$ are manually tuned hyperparameters. We did experiment with other architectures (e.g. RNNs, Transformers instead of the LSTM) but this was the best architecture we found.

4.3. Results

We report F1 scores on tagging in table 5. The best model is the Indic-Transformers BERT, with a distilled version (that is more efficient) coming in a close second. It is surprising that multilingual language models perform much worse than monolingual ones; IndicBERT, for example, is barely better than the baseline. These results are also competitive with F1 scores on English SNACS tagging, which bodes well for future work on multilingual SNACS given the complexity of the Hindi case marker system.

One issue that was prevalent across models was tokenisation errors involving the Devanagari script. The Indic-Transformers models droppped the vowel markers for *u*, *ū*, *e*, *ai*, the *bindu* (nasalisation marker), and the *halant* (vowel-killer) while tokenising. Somehow, they still were the highest-performing models; it is likely that with a fixed tokeniser and retraining they could have been even better.

Nevertheless, these are promising results, especially considering that the complex Hindi case system requires knowledge of verb frame semantics to accurately tag with SNACS.

5. Conclusion

We released an annotated corpus for Hindi of semantic relations encoded by case markers and adpositions, using the multilingual SNACS schema. We presented analysis of the distribution of labels and annotator agreement, explored linguistic issues encountered in annotation that pose problems for SNACS, and ran experiments on automatic sequence tagging for SNACS in Hindi with language models and biLSTM-CRF. We show that this is a feasible computational task and hope that this guides further work on SNACS for other languages, especially for those related to Hindi.

Future work on SNACS could consider multilingual comparisons, building upon work on aligning Korean and English annotations (Hwang et al., 2020) and multilingual tagging as explored in this paper. Particularly, there is ongoing work on SNACS annotation of Gujarati, a langauge closely related to Hindi; multilingual tagging of the two would be an interesting next task. Leveraging SNACS annotations for upstream tasks is also underexplored, despite a growing interesting in the semantic relations encoded in prepositions which have otherwise been understudied in NLP (Elazar et al., 2021). We hope that this corpus will also be useful for future study on semantics-reliant tasks in Hindi.

Acknowledgements

We thank members of the CARMLS group (particularly Maitrey Mehta, Jena Hwang, and Vivek Srikumar) for stimulating discussions on case and adposition semantics, students in NERT, and the three anonymous reviewers for their helpful feedback.

6. Bibliographical References

References

Arora, Aryaman, Meister, Clara, and Cotterell, Ryan (2022). Estimating the entropy of linguistic distributions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. doi: 10.48550/ARXIV.2204.01469.

Arora, Aryaman, Venkateswaran, Nitin, and Schneider, Nathan (2021a). Hindi-Urdu Adposition and Case Supersenses v1.0. *arXiv:2103.01399 [cs.CL]*.

Arora, Aryaman, Venkateswaran, Nitin, and Schneider, Nathan (2021b). SNACS annotation of case markers and adpositions in Hindi. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 454–458. Association for Computational Linguistics, Online.

Begum, Rafiya, Husain, Samar, Bai, Lakshmi, and Sharma, Dipti Misra (2008). Developing verb frames for Hindi. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), Marrakech, Morocco.

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

Bhatt, Rajesh and Anagnostopoulou, Elena (1996). Object shift and specificity: Evidence from ko-phrases in Hindi. *Papers from the main session of CLS*, 32:11–22.

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

Chao, Anne and Shen, Tsung-Jen (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 10(4):429–443.

Conneau, Alexis, Khandelwal, Kartikay, Goyal, Naman, Chaudhary, Vishrav, Wenzek, Guillaume, Guzmán, Francisco, Grave, Edouard, Ott, Myle, Zettlemoyer, Luke, and Stoyanov, Veselin (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics, Online. doi:10.18653/v1/2020.acl-main.747.

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

de Hoop, Helen and Narasimhan, Bhuvana (2005). Differential case-marking in Hindi. In Amberber, Mengistu and de Hoop, Helen, editors, *Competition and Variation in Natural Languages*, Perspectives on Cognitive Science, pages 321–345. Elsevier, Oxford. doi:https://doi.org/10.1016/B978-008044651-6/50015-X.

de Hoop, Helen and Narasimhan, Bhuvana (2009). Ergative case-marking in Hindi. In *Differential subject marking*, pages 63–78. Springer.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota. doi:10.18653/v1/N19-1423.

Elazar, Yanai, Basmov, Victoria, Goldberg, Yoav, and Tsarfaty, Reut (2021). Text-based NP enrichment. *arXiv:2109.12085 [cs]*.

Gardner, Matt, Grus, Joel, Neumann, Mark, Tafjord, Oyvind, Dasigi, Pradeep, Liu, Nelson F., Peters, Matthew, Schmitz, Michael, and Zettlemoyer, Luke (2018). AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6. Association for Computational Linguistics, Melbourne, Australia. doi:10.18653/v1/W18-2501.

Gessler, Luke, Blodgett, Austin, Ledford, Joseph, and Schneider, Nathan (2022). Xposition: An online multilingual database of adpositional semantics. In *Proc. of LREC*. Marseille, France.

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

Hwang, Jena D., Bhatia, Archna, Han, Na-Rae, O'Gorman, Tim, Srikumar, Vivek, and Schneider, Nathan (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. Association for Computational Linguistics, Vancouver, Canada. doi:10.18653/v1/S17-1022.

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

Jain, Kushal, Deshpande, Adwait, Shridhar, Kumar, Laumann, Felix, and Dash, Ayushman (2020). Indic-transformers: An analysis of transformer language models for Indian languages. In *ML-RSA @ NeurIPS 2020*.

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

Joshi, Pratik, Santy, Sebastin, Budhiraja, Amar, Bali, Kalika, and Choudhury, Monojit (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. Association for Computational Linguistics, Online. doi:10.18653/v1/2020.acl-main.560.

Kachru, Yamuna (2006). *Hindi*. Number 12 in London Oriental and African Language Library. John Benjamins Publishing.

Kakwani, Divyanshu, Kunchukuttan, Anoop, Golla, Satish, N.C., Gokul, Bhattacharyya, Avik, Khapra, Mitesh M., and Kumar, Pratyush (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961. Association for Computational Linguistics, Online. doi:10.18653/v1/2020.findings-emnlp.445.

Khanuja, Simran, Bansal, Diksha, Mehtani, Sarvesh, Khosla, Savya, Dey, Atreyee, Gopalan, Balaji, Margam, Dilip Kumar, Aggarwal, Pooja, Nagipogu, Rajiv Teja, Dave, Shachi, et al. (2021). Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Koul, Omkar N. (2008). *Modern Hindi Grammar*. Dunwoody Press.

Kranzlein, Michael, Manning, Emma, Peng, Siyao, Wein, Shira, Arora, Aryaman, and Schneider, Nathan (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. Association for Computational Linguistics, Barcelona, Spain.

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

Liu, Nelson F., Hershcovich, Daniel, Kranzlein, Michael, and Schneider, Nathan (2021). Lexical semantic recognition. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 49–56. Association for Computational Linguistics, Online. doi:10.18653/v1/2021.mwe-1.6.

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

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

Mohanan, Tara (1994). Case OCP: A constraint on word order in Hindi. *Theoretical perspectives on word order in South Asian languages*, 185:216.

Montaut, Annie (2018). The rise of differential object marking in Hindi and related languages. *Studies in Diversity Linguistics*, (19).

Montrul, Silvina, Bhatt, Rakesh, and Girju, Roxana (2015). Differential object marking in Spanish, Hindi, and Romanian as heritage languages. *Language*, pages 564–610.

Montrul, Silvina A, Bhatt, Rakesh M, and Bhatia, Archna (2012). Erosion of case and agreement in Hindi heritage speakers. *Linguistic Approaches to Bilingualism*, 2(2):141–176.

Pal, Riya and Sharma, Dipti (2019). A dataset for semantic role labelling of Hindi-English code-mixed tweets. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 178–188. Association for Computational Linguistics, Florence, Italy. doi: 10.18653/v1/W19-4020.

Paszke, Adam, Gross, Sam, Massa, Francisco, Lerer, Adam, Bradbury, James, Chanan, Gregory, Killeen, Trevor, Lin, Zeming, Gimelshein, Natalia, Antiga, Luca, Desmaison, Alban, Kopf, Andreas, Yang, Edward, DeVito, Zachary, Raison, Martin, Tejani, Alykhan, Chilamkurthy, Sasank, Steiner, Benoit, Fang, Lu, Bai, Junjie, and Chintala, Soumith (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Paul, Soma, Mathur, Prashant, and Kishore, Sushant (2010). Syntactic construct : An aid for translating English nominal compound into Hindi. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 32–38. Association for Computational Linguistics, Los Angeles, California.

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

Prange, Jakob and Schneider, Nathan (2021). Draw mir a sheep: A supersense-based analysis of German case and adposition semantics. *Künstliche Intelligenz*, 35(2).

Qi, Peng, Zhang, Yuhao, Zhang, Yuhui, Bolton, Jason, and Manning, Christopher D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics, Online. doi:10.18653/v1/2020.acl-demos.14.

Ramanathan, Ananthakrishnan, Choudhary, Hansraj, Ghosh, Avishek, and Bhattacharyya, Pushpak (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. Association for Computational Linguistics, Suntec, Singapore.

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

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

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

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

Schneider, Nathan, Hwang, Jena D., Srikumar, Vivek, Prange, Jakob, Blodgett, Austin, Moeller, Sarah R., Stern, Aviram, Bitan, Adi, and Abend, Omri (2018b). Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196. Association for Computational Linguistics, Melbourne, Australia. doi:10.18653/v1/P18-1018.

Schneider, Nathan and Smith, Noah A. (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. Association for Computational Linguistics, Denver, Colorado. doi:10.3115/v1/N15-1177.

Shalev, Adi, Hwang, Jena D., Schneider, Nathan, Srikumar, Vivek, Abend, Omri, and Rappoport, Ari (2019). Preparing SNACS for subjects and objects. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 141–147. Association for Computational Linguistics, Florence, Italy. doi:10.18653/v1/W19-3316.

Spencer, Andrew (2005). Case in Hindi. In *Proceedings of the LFG05 Conference*, pages 429–446. CSLI Publications Stanford, CA.

Vaidya, Ashwini, Choi, Jinho, Palmer, Martha, and Narasimhan, Bhuvana (2011). Analysis of the Hindi Proposition Bank using dependency structure. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 21–29. Association for Computational Linguistics, Portland, Oregon, USA.

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

Wolf, Thomas, Debut, Lysandre, Sanh, Victor, Chaumond, Julien, Delangue, Clement, Moi, Anthony, Cistac, Pierric, Rault, Tim, Louf, Remi, Funtowicz, Morgan, Davison, Joe, Shleifer, Sam, von Platen, Patrick, Ma, Clara, Jernite, Yacine, Plu, Julien, Xu, Canwen, Le Scao, Teven, Gugger, Sylvain, Drame, Mariama, Lhoest, Quentin, and Rush, Alexander (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, Online. doi:10.18653/v1/2020.emnlp-demos.6.