Annotating Sanskrit corpus: adapting IL-POSTS

Girish Nath Jha
Special Center for Sanskrit Studies
Jawaharlal Nehru University, New Delhi
girishjha@gmail.com

Madhav Gopal
Center for Linguistics, SLL&CS,
Jawaharlal Nehru University, New Delhi
mgopalt@gmail.com

Diwakar Mishra
Special Center for Sanskrit Studies
Jawaharlal Nehru University, New Delhi
diwakarmishra@gmail.com

Abstract

In this paper we present an experiment on the use of the hierarchical Indic Languages POS Tagset (IL-POSTS) (Baskaran et al 2008 a&b) , developed by Microsoft Research India (MSRI) for tagging Indian languages, for annotating Sanskrit corpus. Sanskrit is a language with richer morphology and relatively free word-order. The authors have included and excluded certain tags according to the requirements of the Sanskrit data. A revision to the annotation guidelines done for IL-POSTS is also presented. The authors also present an experiment of training the tagger at MSRI and documenting the results.

1. Introduction

Sanskrit, the oldest classical language of India, is also the oldest documented language of the Indo-European family. *ṛgveda* (1500 BCE) is the oldest text of this family contains a sophisticated use of the pre-Pāṇinian variety also called *vaidikī*. Pāṇini variously calls his mother tongue *bhāṣā* or *laukikī*. His grammar has two sets of rules – for *vaidikī* (variety used in the vedas) and for *laukikī* (variety used by the common people). The term 'Sanskrit' (meaning 'refined') is given to the standard form of *laukikī* (current language) which emerged after Pāṇini's grammar *Aṣṭādhyāyī* (AD) (700 BCE) (Jha et al 2007).

Structurally Sanskrit is relatively free word-order and inflected language with amazing capacity to synthesize new sounds and morphemes at the word/sound junctures. Morphologically very elaborate, the rules of Sanskrit grammar are precisely done by Pāṇini. However, these rules are not always easy to completely solve computationally (Jha et al 2009). The viability and usefulness of POS annotation in Sanskrit has been often doubted with the argument that a good morphological analyzer is what would be needed for it. But resolution of ambiguous labels in Sanskrit, as in other languages, cannot be done by morph analyzer alone.

The relatively large computational linguistics community in India does not have a single standard for annotating linguistic data. There are in fact many 'standards'. Five language families with diversity being a norm than exception, the feasibility of a single framework was considered un-attainable until 2008, when Microsoft Research India initiated a collaborative initiative in developing a common framework for Indian languages based on EAGLES guidelines. The framework which resulted is called IL-POSTS (Indic Languages-POS Tag Set) (Baskaran et al 2008 a&b). This framework has since been successfully tried on many Indian languages across many families. The present research is a report on adapting this framework on Sanskrit.

2. POS Tagging in Sanskrit

While POS tagging is not a new research topic, it is, indeed, a new field as far as Sanskrit is concerned. The usefulness of annotated corpora for natural language analyses is well known. Unfortunately, so far, there have been no such annotated corpora available for Sanskrit. An even greater problem has been the lack of training and testing data. The POS information is very important for language processing as it gives significant information about the word and its environment. This is not only true for major grammatical categories (i.e. verb *paṭhati* versus participle *paṭhati*: the verb *paṭhati* expects a noun in nominative but a participle *paṭhati* expects a noun in locative), but also for many other finer distinctions. For example, the declension sub-tags (attributes) and gender-number sub-tags can help distinguish words in different categories (as in *naraḥ gacchanti* the number tag in the verb *gacchanti* demands *naraḥ* to be the nominative plural of the base *nṛ,* and not the nominative plural of the word *nara*). By knowing the POS of a word we can tell which word with which POS label is likely to occur in its vicinity. In linguistic items, such complexity is found across languages.

POS ambiguities in Sanskrit can be enormous. Chandrashekar (2007) has found eight types of ambiguity in the context of POS tag application. In Sanskrit, the nominal base (*prātipadika*) is inflected for multiple information based on the end-character, gender (*liṅga*), *vibhakti* and number (*vacana*) information. The nominal forms (*subanta-*

*rūpa*s) having *vibhakti* markings *-bhyām*, *-bhyas*, *-os* are ambiguous in all endings and genders of a nominal bases (*prātipadika*). Sometimes different nominal bases have similar forms (for e.g., *vibhavaḥ* [1.1] when the *prātipadika* is *vibhava* / *vibhavaḥ*[1.3] when the *prātipadika* is *vibhu*). (Chandrashekar 2007). In a verb, the verb root, along with the optional prefix information like *pada* (*ātmane/ parasmai*), transitivity (*karmatva*), tense (*kāla*), mode (*artha*), voice (*vācya*), person (*puruṣa*) and number (*vacana*) are clubbed together. Adding to the complication, the nominal and verbal bases may be derived bases. Such being the complexity of Sanskrit morphology, there is ample scope for ambiguous word forms.

The complexity of ambiguity in Sanskrit can be demonstrated using an example *bhavati.* It can be a verb or pronoun or a participle. It can still have many more ambiguous forms within the above mentioned categories, if we take inflectional features into consideration as well. Some forms of the first person pronoun *bhavat* in three genders can have similar forms as that of forms in the present participle of the verb root *bhu*. Though actual usage of the *bhavati* in the sense of participle and pronoun in neuter gender is often not seen, but grammaticality of the usage cannot be ruled out.

3. Sanskrit Morphology

In Sanskrit, a syntactic unit is called *pada*. Cordona (1988) posits the formula for Sanskrit sentence (N-En)p…(V-Ev)p. A *pada* can be nominal (*subanta*) or verbal (*tiṅanta*). *Padas* with *sup* (nominal) inflections constitute the NPs (*subanta-pada*), and those with *tiṅ* (verbal) can be called constituting the VPs (*tiṅanta-pada*). In the former, the bases are called *prātipadikas* which undergo sup affixations under specifically formulated conditions of case, gender, number, and also the end-characters of the bases to yield nominal syntactic words. The rules for *subanta padas* are found scattered in AD mostly in chapters 7-1, 7-2, 7-3, 6-1, 6-4. However, these rules have been treated in the *subanta* chapter of *Siddhānta Kaumudī* from rule number 177 to 446. (Jha 2004 b)

The derivational morphology in Sanskrit studies primary forms (*kṛdanta*) and secondary forms (*taddhita*), compounds (*samāsa*), feminine forms (*strī pratyaya*) etc (Subash 2006). These can be inflected for 21 case (7 cases x 3 number) affixes to generate 21 inflected forms.

The verb morphology (*tiṅanta*) is equally complex. Sanskrit has approximately 2014 verb roots including *kaṇḍvādi* according to Pāṇinian *dhātupāṭha* classified in 10 *gaṇas* to undergo peculiar operations (Jha 2004 a), it can also be sub-classified in 12 derivational suffixes. A verb root conjugates for tense, mood, number and person information. Further, these can have *ātmanepadī* and *parasmaipadī* forms in 10 *lakāras* and 3x3 person and number combinations. There are 12 secondary suffixes added to verb roots to create new verb roots. A verb root may have approximately 2190 (tense, aspect, number etc.) morphological forms. Mishra and Jha (2005) have done a rough calculation of all potential verb forms in Sanskrit to be more than 1029,60,000.

4. MSRI hierarchical tagset schema

MSRI in collaboration with linguists and NLP experts has developed a common POS Tagset framework for Indian languages (especially for Indo-Aryan and Dravidian languages) following the hierarchical and decomposable tagset schema similar to that of EAGLES. This framework facilitates the sharing and reusability of scarce resources in Indian languages and ensures cross-linguistic compatibility. The rationale behind concentrating on Dravidian and IA language families has been that of the 22 official languages in India a large majority is associated with these two language families.

The Dravidian and IA language families have very different morpho-syntactic features at every level of linguistic analyses, but they also have a number of typological similarities that facilitate a common framework.

Unlike flat tagsets, a hierarchical tagset exploits the linguistic hierarchy among categories. This implies that instead of having a large number of independent categories, a hierarchical tagset accommodates a small number of categories at the top level, each of which has a number of sub-categories in a tree structure. The associated morpho-syntactic features are packed in the different layers of hierarchy beginning from the major categories in the top and gradually progressing down to cover morpho-syntactic features for making it suitable to any Indian language, thereby keeping the framework a common standard across languages.

The hierarchical tagset requires another feature called ‘decomposability’. It allows different features to be incorporated in a tag by separate sub-strings. Decomposable tags help in better corpus analysis (Leech, 1997) by allowing to search with an underspecified search string.

4.1. The IL-POSTS

This framework has a hierarchy at three levels:

A. *Categories* are the highest level lexical classes. All categories are obligatory, i.e., are generally universal for all languages.

B. *Types* are subclasses of categories and are assumed to be significant sub-classes common to a majority of languages. Some types may also be optional for certain languages.

C. *Attributes* are morpho-syntactic features of types. Attribute tags contain the information like gender (masculine, feminine, neuter), number (singular, dual, and plural), case (nominative, accusative etc.), person (first, second, third) etc. All attributes are optional, though in some cases they may be recommended.

The framework consists of 11 categories (including the punctuations and residual categories) that are recognised as universal categories for all ILs and hence, these are obligatory for any tagset derived from IL-POSTS.

Barring punctuations, all categories have sub-classes called types which can have a number of attributes belonging to each of them. There are 17 attributes defined in the IL-POSTS framework. The attributes can be either binary or multi-valued.

5. Adaptations for Sanskrit

For finalising our tagset, we have used the Sanskrit tagset by Chandrashekar (2007) and the Hindi specific tagset of MSRI. There are some changes in the tagset that we have adapted for Sanskrit data. These are at the category, subcategory, and attribute levels. In the subcategories of noun, we have only common and proper nouns. Verbal nouns have been clubbed under common nouns and spatio-temporal nouns under adverbs. Hindi and Sanskrit nouns do not have similar attributes. The verb has only finite form in Sanskrit and there is no auxiliary. The *upapada ‘sma*’ behaves like an auxiliary but we have put it under particle (*avyaya*). There is no change in the subtypes of pronoun. However, they differ in their attributes. The pronominal takes gender, number, person, case, nominal declension, emphatic, honorificity, and distance attributes. The reflexive and reciprocal carry gender, number, case, and nominal declension attributes. The relative pronoun has gender, number, person, case, and nominal declension. And finally, the Wh-pronoun takes gender, number, person, case, and nominal declension. The subtypes of the demonstrative are same as in Hindi. The adverb category in which a number of Sanskrit indeclinables fall has no attribute like Hindi adverbs. The postposition is not found in Sanskrit. Their role has been replaced by different declensions. In Sanskrit, we have participles (*kṛdanta*) with subtypes as participle proper and participle gerundive. They are marked for gender, number, case, and nominal declension attributes. Under the particle category we have two extra subtypes-negative and emphatic. Particles have no attribute. Punctuation and residual have the same conditions in our tagset.

6. Proposed IL-POSTS for Sanskrit

Following is the tagset we propose for annotating Sanskrit corpus. Using this tagset, we have annotated a corpus of simple Sanskrit text including articles and stories which are online available in our website at http://sanskrit.jnu.ac.in/corpora/annotated/MSRIndic-JNUTagsetTaggedCorpora.txt. Currently, annotation of several Sanskrit story-collections is in progress.

Category	**Type**	**Attributes**
Noun (N)	Common (NC)	gender, number, case, nominal declension
	Proper (NP)	gender, number, case, nominal declension
Verb (V)		pada, number, person, tense\mood, honorificity
Pronoun (P)	Pronominal (PPR)	gender, number, person, case, nominal declension, emphatic, honorificity, distance
	Reflexive (PRF)	gender, number, case, nominal declension
	Reciprocal (PRC)	gender, number, case, nominal declension
	Relative (PRL)	gender, number, person, case, nominal declension
	Wh (PWH)	gender, number, person, case, nominal declension
Nominal Modifier (J)	Adjective (JJ)	gender, number, case, nominal declension, emphatic, negative, honorificity
	Quantifier (JQ)	gender, number, case, nominal declension, numeral, emphatic, negative
Demonstrative (D)	Absolutive (DAB)	gender, number, person, case, nominal declension, distance, honorificity
	Relative (DRL)	gender, number, person, case, nominal declension, distance, honorificity
	Wh- (DWH)	gender, number, person, case, nominal declension, distance, honorificity
Kṛdanta (KD)	Participle (KDP)	gender, number, case, nominal declension
	Gerundive	gender, number, case, nominal

	(KDG)	declension
Particle (C)	Coordinating (CCD)	
	Subordinating (CSB)	
	Gerundive (CGD)	
	Interjection (CIN)	
	Negative (CNG)	
	Emphatic (CEM)	
	Interrogative (CNT)	
	Adverb (CAD)	
	Postposition (upapada) (CPP)	
	Quotative (CQT)	
	Comparative (CCM)	
	Reduplicative (CRD)	
	Other (CX)	
Punctuation (PU)		
Residual (RD)	Foreign word (RDF)	
	Symbol (RDS)	
	Others (RDX)	

Table1. Categories, types and their respective attributes for Sanskrit

No.	**Attributes**	**Values**
1.	Gender (Gen)	Masculine (mas), Feminine (fem), Neuter (neu)
2.	Number (Num)	Singular (sg),Dual (du), Plural (pl)
3.	Person (Per)	First (1), Second (2), Third (3)
4.	Case (Cs)	Nominative (nom), Accusative (acc), Instrumental (ins), Dative (dat), Ablative (abl), Genetive (gen), Locative (loc), vocative (voc),
5.	Nominal declension vibhakti (Vbh)	prathamā (i), dvitīyā (ii), tritīyā (iii), caturthī (iv), pañcamī (v), ṣaṣṭhī (vi), saptamī (vii), vocative (viii)
6.	Tense/Mood (Tns/Mood)	Present (prs), Aorist (aor), Imperfect (imprf), Perfect (prf), Periphrastic Future (phf), General Future (gft), Imperative (imp), Potential (pot), Benedictive (ben), Conditional (cnd
7.	Numeral (Nml)	Ordinal (ord), Cardinal (crd), Non-numeral (nnm)
8.	Distance (Dist)	Proximal (prx), Distal (dst)
9.	Emphatic (Emph)	Yes y, No n
10.	Negative (Neg)	Yes y, No n
11.	Honorificity (Hon)	Yes y, No n
12.	Pada (Pd)	parasmaipada (ppd), ātmanepada (apd)

Table2. Attributes and their values for Sanskrit

We have also used the following common attributes that MSRI Tagset contains:

- Not-applicable (0); when any other value is not applicable to the category or the relevant morpho-syntactic feature is not available.
- Undecided or doubtful (x); when the annotator is not sure about the exact attribute.

7. POS results and current status

The initial experiment of tagging 10 K data of ordinary Sanskrit and subsequent training of the tagger at MSRI is documented below -

Training data: 200 sentences (~5.8K words)
Test Data: 50 Sentences (~1.2K words)

Word Level Acc: 75.35%
Sentence Level Acc: 29.3%

In the similar experimental set up, the accuracy of Sanskrit was much better than Bangla (71%) and lower to Hindi (77%). Currently, we are tagging about 50 K data from two story collections - *Pañcatantra* and *Hitopadeśa*. This task will finish in about two months time. The subsequent training and automatic tagging results will be reported.

8. Conclusion

In this paper we have presented a Sanskrit specific tagset framework for annotating Sanskrit corpus. At experiment level a certain amount of data has been manually tagged and we have revised our tagset again and again. This framework follows the guidelines of the IL-POSTS framework for Indic Languages as much as possible. We have tried our level best to be near to this hierarchical framework. It has also been observed that IL-POSTS framework is adaptable for Sanskrit as well. This Sanskrit Tagset along with the annotation guidelines (that we ourselves have designed for tagging Sanskrit text) and tagged corpus is available on our website: http://sanskrit.jnu.ac.in.

References

AU-KBC tagset. AU-KBC POS tagset for Tamil. Retrieved from http://nrcfosshelpline.in/smedia/images/downloads/Tamil_Tagset-opensource.odt

Baskaran, S. Kalika Bali, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Monojit Choudhury, Girish Nath Jha, Rajendran S., Saravanan K., Sobha L., KVS Subbarao (2008). *A Common Parts-of-Speech Tagset Framework for Indian Languages,* LREC 2008 - 6th Language Resources and Evaluation Conference, May 26-June1, 2008, Marrakech, Morocco

Baskaran S. Kalika Bali, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Monojit Choudhury, Girish Nath Jha, Rajendran S., Saravanan K., Sobha L., KVS Subbarao (2008). *Designing a Common POS-Tagset Framework for Indian Languages.* The 6thWorkshop on Asian Language Resources. January, 2008, Hyderabad.

Cardona, George, 1988. *Pāṇini: His work and its traditions*, Motilal Banarasidass, Delhi

Chandrashekar R., 2007. *POS Tagger for Sanskrit*, Ph.D. thesis, Jawaharlal Nehru University

Cloeren, J., 1999. *Tagsets. In Syntactic Wordclass Tagging*, ed. Hans van Halteren, Dordrecht.: Kluwer Academic

Jha Girish Nath, 2004. *Generating nominal inflectional morphology in Sanskrit*, SIMPLE 04, IIT-Kharagpur Lecture Compendium, Shyama Printing Works, Kharagpur, WB

Jha Girish Nath, Sobha L, Diwakar Mishra, Surjit K Singh, Praveen Pralayankar, 2007. *Sanskrit Anaphors*, Johansson, C. (Ed.) Proceedings of the Second Workshop on Anaphora Resolution (2008), ISSN 1736-6305 Vol. 2, Cambridge Scholars Publishing, 2007

Jha Girish Nath, Mishra Sudhir, 2009. *Semantic processing in Panini's karaka system*, Lecture Notes in Computer Science series 5402 Springer Berlin / Heidelberg, Gérard P. Huet, Amba P. Kulkarni, Peter Scharf (eds)

Greene, B.B. and Rubin, G.M.,1981. *Automatic grammatical tagging of English*. Providence, R.I.: Department of Linguistics, Brown University.

Hardie, A., 2004. *The Computational Analysis of Morphosyntactic Categories in Urdu*. PhD Thesis submitted to Lancaster University.

IIIT-Tagset. *A Parts-of-Speech tagset for Indian Languages*. Retrieved from http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

Huet Gerard, *The Sanskrit Heritage Site*, http://sanskrit.inria.fr/

Kale, M.R., 1995. *A Higher Sanskrit Grammar*, MLBD Publishers, New Delhi

Leech, G and Wilson, A., 1996. *Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Report EAG-TCWG-MAC/R.

Leech, G and Wilson, A., 1999. *Standards for Tag-sets*. In Syntactic Wordclass Tagging, ed. Hans van Halteren, Dordrecht: Kluwer Academic.

Leech, G., 1997. *Grammatical Tagging*. In Corpus Annotation: Linguistic Information for Computer Text Corpora, ed: Garsire, Leech, and McEnery, London: Longman.

Mishra Sudhir, Jha, Girish Nath, 2005. *Identifying verb inflections in Sanskrit morphology*, proceedings of SIMPLE'04, IIT Kharagpur

NLPAI Contest- 2006, retrieved from http://ltrs.iiit.ac.in/nlpai_cntest06

Santorini, B., 1990. *Part-of-speech tagging guidelines for the Penn Treebank Project*. Technical report MS-CIS-90-47, Dept. Of Computer and Information Science, University of Pennsylvania.

Subash, 2006. *Sanskrit Subanta Recognizer and Analyzer*, M.Phil dissertation submitted to Jawaharlal Nehru University, New Delhi