APPLICATION OF BIS POS TAGSET FOR SANSKRIT: CASE OF VERBS AND PARTICLES

Madhav Gopal

Anil Pratap Giri

Girish Nath Jha

Jawaharlal Nehru University, New Delhi

In: National Seminar On POS Annotation for Indian Languages: Issues & Perspectives, 12-13th Dec, 2011, CIIL, Mysore

POS TAGGING

- POS tagging (or morphosyntactic tagging) is the process of assigning to each word in a running text a label which indicates the status of that word within some system of categorizing the words of that language according to their morphological and/or syntactic properties (Hardie, 2003).

It is useful in:

- word sense disambiguation
- developing pos taggers
- chunking
- information retrieval
- machine translation
- parsing

SANSKRIT POS TAGGING

- **Linguistic Nature of Sanskrit:**

- Rich inflectional, derivational morphology
- Various strategies for encoding same information
- Highly synthetic nature of the language
- Complex orthographic system
- Irregularity of punctuation marks

- **Varying number of grammatical categories in the tradition:**

- Indra School: 1- अर्थः पदम्
- Pāṇini: 2- सुबन्तं तिङन्तं पदम्
- Jagadish: 3- प्रकृति, प्रत्यय, निपात
- Yāska: 4- नाम, आख्यात, उपसर्ग, निपात
- New grammarians: 5- नाम, आख्यात, उपसर्ग, निपात, कर्मप्रवचनीय

CONT....

- In Sanskrit we tag a *pada*, a linguistic unit usable in a sentence.
- Due to complex orthographic system of the language, and sandhi operations, sometimes two or more *padas* are concatenated and they seem to be one word, which they are not. These are cases of *anitya sandhi* and must be resolved first. To tag a sequence of words with *anitya sandhi* is impossible.

- ***nitya* (mandatory) vs. *anitya* (optional) sandhi:**

संहितैकपदे नित्या नित्या धातूपसर्गयोः।
नित्या समासे वाक्ये तु सा विवक्षामपक्षते॥

(after sandhi splitting)

संहिता एकपदे नित्या नित्या धातूपसर्गयोः।
नित्या समासे वाक्ये तु सा विवक्षाम् अपक्षते॥

THE BIS POS TAGSET

- This tagset is a national standard tagset for Indian languages that has been recently designed by the Bureau of Indian Standards (BIS henceforth) committee.
- The BIS scheme is comprehensive and extensible and can spawn tagsets for Indian languages based on individual applications.
- This tagset has 11 categories at the top level. The categories at the top level have further subtype level 1 and subtype level 2.
- The standard which has been followed in this tagset takes care of the linguistic richness of Indian languages.
- This is a hierarchical tagset and allows annotation of major categories. Thus, it reduces the cognitive load of human annotator.
- For morphological analysis it will take help from Morphological Analyzer, so morpho-syntactic features are not included in the tags.

SANSKRIT VERBS

- Sanskrit verbs are generally classified in three categories: *parasmaipada*, *ātmanepada* and *ubhayapada*.
- The *parasmaipada* form denotes that the fruit of the action goes to someone different other than the agent whereas the *ātmanepada* form denotes the fruit of the action goes to the agent herself.
- They can again be classified into primary and derivative verbs depending on the type of verbal root.
- Sanskrit verbs encode voice, tense/mood, person and number features.
- They allow prefixation and suffixation and the resultant of these operations remain a *pada*.

THE BIS SCHEME FOR VERB TAGGING

4	Verb		V	V
4.1	Main		VM	V_VM
04/01/01		Finite	VF	V_VM_VF
4/01/02		Non-finite	VNF	V_VM_VNF
04/01/03		Infinitive	VINF	V_VM_VINF
04/01/04		Gerund	VNG	V_VM_VNG
4.2	Auxiliary		VAUX	V_VAUX

TAGGING SANSKRIT VERBS

- **Finite (VF)**
- All the conjugations of the *dhātus* are finite verbs (VF). However, when some of these forms will be used to express the aspectual meaning of the preceding *kṛdanta* will be tagged as auxiliary, as is stated above. In addition, *kta* and *ktavat pratyayāntas* will also be tagged as VF when they are not followed by an auxiliary. As we do not have a separate tag for gerundives (like *kāryam*, *karaṇīyam*, *kartavyam*), VF tag could be applied for them as well.
- मोहनः/NNP हैदराबादम्/NNP **गतवान्/VF** ।/PUNC सः/PRP मम/PRP भ्राता/NN अस्ति/VF ।/PUNC सुषमा/NNP विशाखापत्तनम्/NNP **गच्छति/VF** ।/PUNC

CNTD…

- **Non-finite (VNF)**
  - *kta* and *ktavat pratyayāntas* (these are generally described as participles in literature) will be tagged as verb non-finite (VNF) when followed by an auxiliary and other *kṛidantas* like *śatṛ*, *śānac* and *kānac* will also get the same tag.
  - कल्पना/NNP प्रयागम्/NNP **गच्छन्ती/VNF** तेन/PRP सह/PSP वार्ताम्/NN करिष्यति/VF ।//PUNC अधुना/RB सा/PRP सिंगापुरम्/NNP गता/VNF अस्ति/VAUX ।/PUNC
- **Infinite (VINF)**
  - Sanskrit infinitives are different from other Indian languages and English. They correspond to the infinitive of purpose in English. They are formed by adding *tumun* suffix in the verb root. Only *tumun pratyayāntas* will be tagged as VINF.
  - सा/PRP जयपुरम्/NNP अपि/RPD **गन्तुम्/VINF** इच्छति/VF ।/PUNC

CNTD…

- **Gerund (VNG)**
- In the literature *ktvānta* and *lyabanta* forms are described as gerund. So, these kinds of constructions will be labeled with the gerund (VNG) tag.
- कल्पना/NNP गोरखपुरम्/NNP **गत्वा/VNG** प्रयागम्/NNP गमिष्यति/VF ।/PUNC तत्र/RB च/CCD स्वकीयाम्/PRF मातरम्/NN **आदाय/VNG** गङ्गास्नानम्/NN करिष्यति/VF ।/PUNC

CNTD…

- **Auxiliary (VAUX)**

  - In the language some *tiṅantas* (like verbal inflections of *as*, *ās*, *sthā*, *kṛ*, and *bhū* only) that follow a *kṛdanta* to express its (*kṛdanta's*) aspectual meaning, will be tagged with Auxiliary label and the indeclinable *sma* will also get the same tag when follows a verb in present tense and modifies the meaning of the associated verb.

  - ततः/NST च/CCD पिङ्गलकः/NNP सञ्जीवकेन/NNP सह/PSP सुभाषितगोष्ठीसुखम्/NN अनुभवन्/VNF **आस्ते/VAUX** ।/PUNC तस्मिन्/DMD वने/NN भासुरकः/NNP नाम/JJ सिंहः/NN प्रतिवसति/VF **स्म/VAUX** ।/PUNC सः/PRP अधुना/RB सिंगापुरम्/NNP गतः/VNF **अस्ति/VAUX** ।/PUNC

THE BIS SCHEME FOR PARTICLE TAGGING

9	Particles			RP	RP
9.1		Default		RPD	RP_RPD
9.2		Classifier		CL	RP_CL
9.3		Interjection		INJ	RP_INJ
9.4		Intensifier		INTF	RP_INTF
9.5		Negation		NEG	RP_NEG

SANSKRIT PARTICLES

- **Default Particle (RPD)**

- In the current system this would be applied for all avyayas which don't have specific tag in this framework. This will include the *avyaya* types सादृश्यादि, अवधारणम्, and प्रश्नार्थक.

- **अथ/RPD** किम्/PRQ करणीयम्/VF ?/PUNC सुकुमारा/JJ **खलु/RPD** इयम्/PRP ?/PUNC अपि/RPD गच्छति/VF सः/PRP ?/PUNC आम्/INJ ,/PUNC सः/PRP एव/RPD गन्तुम्/VINF शक्नोति/VF ।/PUNC

- **Classifier Particle (CL)**
- This tag is not applicable for Sanskrit.

CNTD…

- **Interjection (INJ)**
  - Words that express emotion are interjections, and also the particles which we use for getting the attention of people, e.g., बत, अहो, हा, धिक्, स्वधा, हे, भो etc.

  - **भो/INJ** बालकाः/NN !/PUNC यूयम्/PRP किम्/PRQ कुरुथ/VF ?/PUNC

- **Intensifier (INTF)**
  - Adverbial elements with an intensifying role are intensifiers. They could be both, either positive or negative. भृशम्, पूर्णतया, न्यूनतया, न्यूनातिन्यूनम् etc. will fall in this category.

  - तम्/PRP अवेक्ष्य/VNG रुरोद/VF सा/PRP **भृशम्/INTF** ।/PUNC

CNTD…

- **Negation (NEG)**

◉ The indeclinables which are used for expressing negation are treated under this category.

◉ चिन्ता/NN **मा/NEG** करोतु/VF ।/PUNC सः/PRP भवन्तम्/PRP न/NEG ताडयिष्यति/VF ।/PUNC

WHAT ABOUT OTHER *AVYAYAS*?

- Some Sanskrit *avyayas* function as adverbs, so they are tagged as adverb, a separate category in this scheme.
- शनैः/RB शनैः/ECH अग्रे/NST चलामः/VF ।/PUNC
- Some *avyayas* behave like ambiposition (*upapada*). In this framework they belong to the category of Postposition.
- दुर्गम्/NN अभितः/PSP परिखा/NN अस्ति/VF ।/PUNC
- Some *avyayas* function as conjunction/disjunction, they have been put in the separate category of Conjunction.
- नायकः/NN खलनायकः/NN च/CCD सहरूपेण/PSP गच्छन्ति/VF ।/PUNC
- रामः/NNP अकथयत्/VF यत्/CCS सः/PRP आपणम्/NN गमिष्यति/VF ।/PUNC
- Quotative is also under Conjunction:
- "/PUNC सर्वे/PRP भवन्तु/VF सुखिनः/NN "/PUNC इति/UT केन/PRQ उक्तम्/VF ?/PUNC

CONCLUSION

- This scheme captures appropriate linguistic information, and also ensures the sharing, interchangeability and reusability of linguistic resources. The Sanskrit specific tagsets available so far (barring IL-POSTS) are not compatible with other Indian languages and with the exception of the IL-POSTS, all other tagsets are flat and brittle and do not capture the various linguistic information.
- This initiative, we hope, will enrich Indian NLP and will eliminate the language barriers between different linguistic communities not only in India but across the world. The uniformity in tagging all Indian languages will help in identifying linguistic differences and similarities among Indian languages, and thus facilitate other NLP/linguistic researches.

CNTD….

- Moreover, the corpus annotated with this tagset would be more useful as it is tagged by a standard tagset or paradigm. This will ensure the maximal use and sharing of the tagged data. The initiative for tagging Indian languages with the present standard tagset is a promising effort in this direction with the hope that all Indian language corpora annotation programmes will follow these linguistic standards for enriching their linguistic resources. Thus, Indian NLP may grow faster!

REFERENCES

- Gopal, Madhav and Jha, Girish N.: Tagging Sanskrit Corpus Using BIS POS Tagset. In: Singh, C., Lehal, G.S., Sengupta, J., Sharma, D.V., and Goyal, V. (eds.) Proceedings of the International Conference, ICISIL 2011, Patiala, India, March 9-11, 2011, CCIS 139 pp. 191-194, Heidelberg: Springer.
- Chandrashekar, R.: Parts-of-Speech Tagging For Sanskrit. Ph.D. thesis submitted to JNU, New Delhi (2007)
- Gopal, Madhav, Mishra, Diwakar and Singh, Priyanka Devi.: Evaluating Tagsets for Sanskrit. In: Jha, Girish Nath (ed.) Proceedings of the Fourth International Sanskrit Computational Linguistics Symposium, Dec.10-12, 2010, Heidelberg: Springer.
- IIIT-Tagset. A Parts-of-Speech tagset for Indian Languages. http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf
- Jha, Girish Nath, Gopal, Madhav, Mishra, Diwakar.: Annotating Sanskrit Corpus: adapting IL-POSTS. In: Z. Vetulani (ed.) Proceedings of the 4th Language and Technology Conference: Human Language Technologies as a challenge for Computer Science and Linguistics, pp. 467-471 (2009)

Thank You for Your Attention!

mgopalt@gmail.com
apgiri.san@pondiuni.edu.in
girishjha@gmail.com