Guidelines for POS for Sanskrit (Developed Under Sanskrit Consortium Project)

Version 1.6 dated 25.02.2010; same as Version 1.5 with Eng abbr for the Tags

1 Introduction

The significance of Sanskrit Language Processing in the context of Indian Language Processing is well known, for, Sanskrit has great influence on other languages being the mother to all other Indian Languages and it has a well formed grammar formalism which narrates the structure of Sanskrit. For the same reason Sanskrit was claimed to be the most suitable language for computers. It would not be improper to examine the same by developing Language processing tools for Sanskrit using its grammar rules. Annotated corpus of Sanskrit serves as an important tool for investigators of natural language processing, speech recognition and other related areas. It proves to be a basic building block for constructing statistical models for automatic processing of natural languages.

Many such corpora are available for languages across the world and have proved to be a useful step towards natural language processing. Coming to the scenario for Indian languages that too for Sanskrit, not much work has been carried out on the front of automatic processing of Sanskrit. The main bottleneck being unavailability of an annotated corpora, large enough to experiment statistical algorithms, and also to evaluate the perfomance of rule based algorithms.

Annotation of corpus involves various levels viz, part of speech, phrase/clause level, dependency level, etc. Part of speech tagging forms the basic step towards building an annotated corpus. Chunking can form the next level of tagging. In the context of Sanskrit, Samasa tagging is also an additional work.

A consortium has been constituted for developing tools for Sanskrit Language and a MT system for Sanskrit-Hindi initially for Children literature and Scientific fields like Ayurveda. As part of the project, standard tagsets are being developed for annotation of Sanskrit Corpus at various levels.

The issues related to defining standards for Sandhi, POS, Samasa and Karaka level tagging schemes were discussed by scholars from various Indian institutes by way of holding meetings etc. and some standards have been arrived at.

Separate guidelines have been written for each tagging scheme with an exhaustive examples and problems with solutions.

The present document discusses about the POS tagging for Sanskrit.

2 Objective

A series of meetings were held by the consortium members to arrive at standard tagging scheme for annotating Sanskrit Language texts and come up with the POS tags which are exhaustive for the task of annotation. The present document gives a detailed description of the tags which have been defined for the tagging schemes and elaborates the motivations behind the selection of these tags. The document also discusses various issues that were addressed while preparing tagsets and how they have been resolved.

3 Some Assumptions

- During the workshop it was decided to base the discussion and decisions about various tags on the following basic assumptions which everybody agreed on :

  1. The tags should be based on Shastric guidelines.
  2. They should be comprehensive/ complete.
  3. Tags should be in Sanskrit terms used in Shastras.
  4. They should be simple ; Maintaining simplicity is important for the following two reasons :

     (a) Ease of Learning / tagging

     (b) Consistency in annotation

- Another important point which was discussed and agreed upon was that POS tagging is NOT a replacement for morph analyser. A ‘word’ in a text carries the following linguistic knowledge.

  1. grammatical category and
  2. grammatical features such as gender, number, person etc.

  The POS tag should be based on the ‘category’ of the word and the features can be acquired from the morph analyser.

- Another noteworthy point to be noted by the annotators is that it is more important to see whether the tag is easy from the aspect of machine learning within the range of adjacent words, than to keep all information possible.

4 Necessity of POS tagger

Thus, to begin with, it has been decided to adopt Shastric lines which shed light on this issue.

It is noteworthy here that Sanskrit language has been analysed by Shastric Scholars with out the stage called "POS". The whole शाब्दबोध *process which involves many steps, does not stress upon POS analysis. Hence there was a feeling among Sanskrit Scholars that no POS tagging is required for Sanskrit Languagae analysis.*

To understand the role of POS tagger in the analysis of Sanskrit texts let us look at the process of Śābdabodha.

The process of Śābdabodha may be thought of as consisting of following steps:

- Do Pada viccheda
- Analyse each split word at morphological level
- Do the samāsa(compound) analysis wherever necessary
- Decide the main verb
- Looking at the vibhakti, prayoga and lakāra information do the anvaya.
- While doing the anvaya, also decide the viśeṣya-viśeṣaṇa bhāva wherever applicable.

In this process, the important role played by a human being is in deciding the main verb and doing anvaya.

Consider an example: रामः वनम् गच्छति।

In this sentence, रामः (rāmaḥ) can be either a noun or a verb. Similarly गच्छति (gacchati) can be either a derived noun (derived by adding 7^{th} vibhakti to शतृ (Ṣatṛ) pratyayānta gacchat) or a finite verb form of the verb gam.

A human being, based on his/her knowledge about the world, takes a decision about which of these analysis is correct. This decision helps him to know which is the main verb and accordingly based on the आकाङ्क्षा (ākāṅkṣā) of the verb, further analysis is carried out.

The role of POS tagger is to help the machine

- rule out some of the possibilities, thereby reducing the ambiguity at morpho-syntactic level, and also the search space for a parser,
- help in disambiguity across the category,
- provide likely POS tags for the words where morph analyser has failed.

5 Issues in POS Tag Set Design

This section deals with some of the issues related to any POS tagger and the policy that we have adopted to deal with each of these issues for our purpose. The first step towards developing POS annotated corpus is to come up with an appropriate tags. The major issues that need to be resolved at this stage are :

- Fineness vs Coarseness in linguistic analysis
- Syntactic Function vs lexical category
- New tags vs tags close to existing ILMT/JNU tags

5.1 Fineness vs Coarseness

An issue which always comes up while deciding tags for the annotation task is whether the tags should capture 'fine grained' linguistic knowledge or keep it 'coarse'. In other words, a decision has to be taken whether or not the tags will account for finer distinctions of the parts of speech features. For example, it has to be decided if plurality, gender and other such information will be marked distinctly or only the lexical category of a given word should be marked.

It was decided to come up with a set of tags which avoids 'finer' distinctions. The motivation behind this is to have less number of tags since less number of tags lead to efficient machine learning. Further, accuracy of manual tagging is higher when the number of tags is less.

However, an issue of general concern is that in an effort to reduce the number of tags we should not miss out on crucial information related to grammatical and other relevant linguistic knowledge which is encoded in a word. If tags are too coarse, some crucial information for further processing might be missed out. As mentioned above, primarily the required knowledge for a given lexical item is its grammatical category, the features specifying its grammatical information and any other information suffixed into it.

In Sanskrit each word by itself is a bundle of linguistic information. Morph analyser provides all the knowledge that is contained

in a word. It was decided that any linguistic knowledge that can be acquired from any other source (such as morph analyser) need not be incorporated in the POS.

As mentioned above, POS tagger is not a replacement for morph analyser. In fact, features from morph analyser can be used for enhancing the performance of a POS tagger. The additional knowledge of a POS given by a POS tagger can be used to disambiguate the multiple answers provided by a morph analyser.

On the other hand, we agree that too coarse an analysis is not of much use. Essentially, we need to strike a balance between fineness and coarseness. The analysis should not be so fine as to hamper machine learning and also should not be so coarse as to miss out important information. It is also felt that fine distinctions are not relevant for many of the applications(like sentence level parsing, dependency marking, etc.) for which the tagger may be used in future.

However, it is well understood that plurality and other such information is crucial if the POS tagged corpora is used for any application which needs the agreement information. In case such information is needed at a later stage, the same tag set can be extended to encompass information such as plurality etc as well. This can be done by providing certain heuristics or linguistic rules.

5.2 Syntactic Function vs Lexical Category

A word belonging to a particular lexical category may function differently in a given context. For example, the lexical category of कृष्ण in Sanskrit is a noun. However, functionally,कृष्ण is used as an adjective many a times. It is common in Sanskrit any noun could be a विशेषण if it's content is Quality, action or type. There is no adjective category in Sanskrit.

Such cases require a decision on whether to tag a word according to its lexical category or by its syntactic category. Since the word in a context has syntactic relevance, it appears natural to tag it based on its syntactic information. However, such a decision may lead to further complications.

In AnnCorra, the syntactic function of a word is not considered for POS tagging. Since the word is always tagged according to its lexical category there is consistency in tagging. This reduces confusion involved in manual tagging. Also the machine is able to establish a word-tag relation which leads to efficient machine learning.

Unlike modern Indian languages, where suntax distinguishes between an adjective and a noun, Sanskrit does not have any distinction at syntactic level. Hence, after detailed discussion, it was decided to mark विशेषण as a subcategory of Noun etc., in the cases wherever applicable. Later this information can be used for further analysis.

5.3 New Tags vs Tags from a Standard Tagger

Another point that was considered while deciding the tags was whether to come up with a totally new tag set or take any other standard tagger as a reference and make modifications in it according to the objective of the new tagger. It was felt that the later option is often better because the tag names which are assigned by an existing tagger may be familiar to the users and thus can be easier to adopt for a new language rather than a totally new one. It saves time in getting familiar to the new tags and then work on it as in the case ILMT. But, later it was found that Sanskrit scholars find it very difficult to use unfamiliar tags like JJ etc.

The POS tagset for Sanskrit was first developed at JNU[1]. This tagset represents the morph analysis as well, and thus is more fine grained. While drawing a lot from this existing tagset, the morphological information being available at other modules, it was decided to keep the POS tagset coarse.

6 Basic Parts Of Speech

Having established the necessity of the POS tagger for Sanskrit, now let us see, what are the basic POS tags in Sanskrit. It was felt that since “यास्क ” a great grammarian has shown the word categories in sentences, it is not against the shastric tradition. Hence it was resolved to have POS tagging stage which is based on “यास्क’s” categories.

A pada in Sanskrit is defined as ‘सुप् -तिङ् -अन्तम् पदम्’.
Thus there are basic two classes or Parts of Speech viz. सुबन्त and तिङन्त (or नाम and आख्यात).

Yaska classifies the words in 4 different categories:

- नाम (noun)
- आख्यात (verb)
- उपसर्ग (pre-position), and
- निपात (indeclinable)

We discuss below the importance and (hence) the need to include each of these categories in the POS tagger.

[1] by Dr. R Chandrashekhar under the guidance of Dr. Girish Nath Jha

The distinction between नाम (noun) and आख्यात (verb) is useful in itself, since it helps us further in parsing. Further in Vedic literature, उपसर्ग (prepositions) is written separately which may float anywhere around. Identifying and marking them will help in grouping them with the verb. The निपात (indeclinable) do not have any कारक role. Hence it is desirable to mark them in order to reduce the search space of the parser.

Thus Yaska's classification may be taken as a starting point and also as a coarse tagset to tag the Sanskrit words. However, these tags in themself are not sufficient. We need further sub-classification. We suggest below some more tags and also justify the necessity of these tas.

6.1 नाम (noun)

Major concern in case of नाम is whether a particular नाम is a विशेषण or a विशेष्य. Also in case of translation it matters the most to know whether the नाम (noun) is a नामसंज्ञा (Proper name) or not. In case of नामसंज्ञा, the name should not be translated.

Similarly the knowledge of whether a word is सर्वनाम (Pronoun) or a नाम (noun) is essential in order to find its reference. Of course the list of सर्वनाम is lexicon driven, and is a closed list. The सर्वनाम can further be either a विशेषण (modifiers) or a विशेष्य (modified). In case of विशेषण, it need to be chunked together with its विशेष्य, and will not get any kāraka role. The साम्बन्धिक सर्वनाम marks the inter-sentential relations.
Finally the संख्या also need to be classified as

- संख्यावाचक
- संख्येयवाचकम्, and
- संख्यापूरकवाचकम्.

The संख्यावाचक are always in neuter gender and संख्येयवाचक take the gender of the following nouns. Hence there is a necessity of two classes. The संख्येयवाचक can never be विशेष्य. They are always विशेषण, and hence will not get any कारक role but will always get combined with their विशेष्य. The संख्यापूरकवाचक is always a विशेषण and hence it also gets combined with its विशेष्य.

Taking into account all these aspects, we decided to subclassify the नाम (nouns) into following sub-categories

- नाम
- नामविशेषणम्
- नामसंज्ञा

- सर्वनाम
- सर्वनामविशेषणम्
- साम्बन्धिक सर्वनाम
- कृदन्त नाम विशेषणम्
- संख्यावाचकम्
- संख्येयवाचकम्
- संख्यापूरकवाचकम्

6.2 कृदन्त

There is a need to mark the **कृदन्त**, since a **कृदन्त**, though at surface level acts either as an avyaya or a noun, it has its own **आकाङ्क्षा** which needs to be fulfilled. Further, depending on its suffix, it may be further classified as a **कृदन्त -अव्यय** or a **कृदन्त - विशेषण**. A **कृदन्त - विशेषण** will get grouped with its **विशेष्य** and will not have any **कारक** role to play. A **कृदन्त -अव्यय** on the other hand will mark a relation with the finite verb.

6.3 Avyayas

Avyaya is a bag of words. It consistes of several types of indeclinables, and their functions are different. Therefore it is necessary to classify them further.

6.3.1 समुच्चय and प्रश्नार्थक

For example, consider the avyaya '**अपि**'. Position of this avayaya in the sentence decides whether the sentence is declarative or interrogative, as in the following cases:

अपि सः गच्छति ?
सः अपि तेन सह गच्छति.

So there is a necessity to mark the avyaya as

- समुच्चय
- प्रश्नार्थक

6.3.2 भावसूचक and सम्बोधनम्

Same avyaya can be used as a उद्गारवाचक and for सम्बोधनम्. For example:
हा हन्त लक्ष्मणः पतितः।
हे राम, गच्छ।
भो राम, अत्र आगच्छ।
If a POS tagger can mark the avyaya as one of the two, it helps in establishing the relations further.

6.3.3 उद्धरण

उद्धरण words act as vibhkati markers for the vākya karma. But some of the words used for उद्धरण are ambiguous, such as इति. रावणो हतम् इति श्रुतम्
मम मित्रः आगतवान् इति अहम् गृहम् गच्छामि
In the first sentence 'इति' is used as an उद्धरण, whereas in the second sentence it marks the relations between two sentences.

6.3.4 द्विरुक्ति

The marking of dviruktis is useful in grouping the words together, as they give a single meaning.

Example: स्मारम् स्मारम् तिष्ठति.

6.3.5 निषेधार्थक

The निषेधार्थक avyayas change the meaning of the verbs and hence they should be grouped together with the verbs. Marking these avyayas help in grouping them with the verbs.
Examples: रामो नगरीम् न आगच्छति।

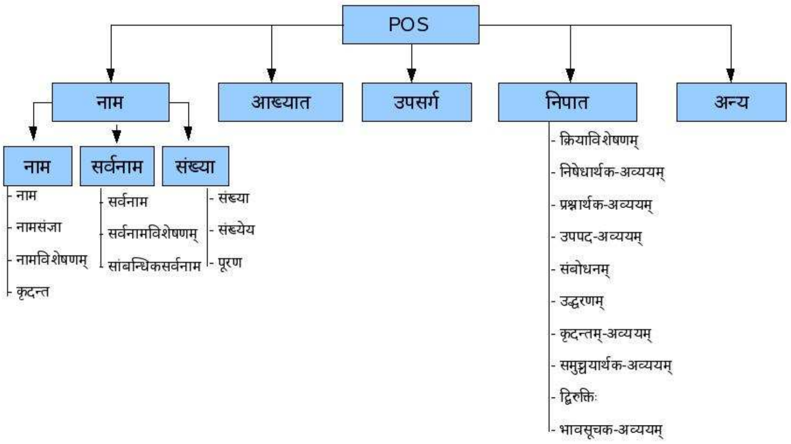
6.3.6 उपपद अव्यय

The upapada avyayas should be grouped with the preceeding nouns leading to a meaningful unit to be processed. Therefore they need to be grouped.
Examples:
रामेण सह
ग्रामम् परितः

6.3.7 क्रिया विशेषण

Finally the क्रियाविशेषण should be marked since it just gives more information about the manner in which the activity is carried out, or the special property of the resultant of the activity.

सः मन्दं गच्छति।
सः ओदनं स्तोकं पचति।



Thus following is the proposed classification of avyayas:

- क्रियाविशेषणम्
- उपपद अव्ययम्
- निषेधार्थक अव्ययम्
- समुच्चयात्मक अव्ययम्
- प्रश्नार्थक अव्ययम्
- कृदन्त अव्ययम्
- उद्गारसूचक अव्ययम्
- सम्बोधनम्
- उद्धरणम्
- द्विरुक्ति

The hierachical tagset is shown in Figure 1.

Table 1 gives the equivalent TAG names of existing tagsets.

Description	*SKT ABBR*	*SKT Tag name*	*JNU Tag*	*ILMT Tag*	*Microsoft Tag*
नामपदम्	(नाम)	N	N	NN	
नामसंज्ञा	(नाम सं)	NS	NA	NNP	
नाम-विशेषणम्	(नाम-वि)	NV	NVI	JJ	
कृदन्तनाम-विशेषणम्	(कृ-ना-वि)	KNV	KV	-	
सर्वनाम	(सर्व)	SN	SN	PRP	
सर्वनाम - विशेषणम्	(सर्व-वि)	SNV	–	DEM	
साम्बन्धिक - सर्वनाम	(सर्व-साम्)	SNS	SNS	-	
संख्यावाचक	(सं)	SAM	SAM	-	
संख्येयवाचक	(संय)	SAMY	SAMY	QC	
संख्यापूरणवाचक	(सं-पू)	SAMP	–	QO	
क्रियापद	(क्रि)	KP	several tags	VM	
उपसर्ग	(उप)	UP	-	-	
निषेधार्थक - अव्ययम्	(निषे-अ)	AVN	AVN	NEG	
प्रश्नार्थक - अव्ययम्	(प्र)	AVP	AVP	WQ	
उपपद - अव्ययम्	(उप-अ)	AVUP		-	
सम्बोधनम्	(सम्बो)	SAMB	-	-	
उद्धरण	(उद्ध)	AVU	-	-	
कृदन्त - अव्ययम्	(कृ)	K	AK	-	
क्रियाविशेषण - अव्ययम्	(क्रि-वि)	AKV		-	
भावसूचक - अव्ययम्	(भाव-अ)	AB	UD	-	
समुच्चायक - अव्ययम्	(समु-अ)	AVC	AVC	CC	
द्विरुक्ति	(द्वि)	DVI	-	-	
विकल्पार्थक - अव्ययम्	(विकल्प - अ)	AV	AVD		

Table 1: Version 1.1

7 Modifications: Version 1.2

The POS tag set shown in Table 1 is modified after the meeting on 17-18th December, 2008.
Modified tagset is shown in the Table 2.
The modifications are:

- A new tag कृदन्तनाम is added.

- उपसर्ग tag is needed only in case of Vedic literature, and its presence in this tagset, may cretae confusion. Hence to avoid the confusion, the tag has been removed.

- सम्बोधनम् is changed to सम्बोधन-सूचक, to avoid the confusion in tagging. This tag is to mark the words such as 'हे', 'भो' etc as सम्बोधन सूचक, and not to mark the words 'भगवन्' as सम्बोधन. भगवन् will be marked as a नामपदम् only.

- विकल्पार्थक is removed. It is now merged with the समुच्चायक.

- समुच्चायक is changed to yojaka to incorporate all the योजकs. योजकs are further classified into two: a पद योजक and a वाक्य योजक.

- A new tag is introduced to handle सादृश्यादि. This contains all the remaining avyayas such as ‘इव’, ‘एव’, ‘अपि’, etc.

- To mark the words from other languages a new tag ‘अन्य भाषा’ is introduced.

- The punctuation markes will be marked as they are.

- In case of doubt, it was decided to put a ‘?’, which can then be resolved later.

- In case of confusion, or rare cases where it is not clear which tag to assign a tag ‘अज्ञात’ is assigned, but this should be used very cautiously, only in rare cases.

Description	*SKT ABBR*	*JNU Tag*	*ILMT Tag*
नामपदम्	(नाम)	NP	NN
नामसंज्ञा	(नाम_सं)	NA	NNP
नाम-विशेषणम्	(नाम_वि)	NVI	JJ
कृदन्तनाम	(कृ_नाम)	several tags	VM
कृदन्तनाम-विशेषणम्	(कृ_नाम-वि)	several tags	VM
सर्वनाम	(सर्व)	SN	PRP
सर्वनाम - विशेषणम्	(सर्व_वि)	SNN	DEM
संख्यावाचक	(संख्या)	SAM	QC
संख्येयवाचक	(संख्येय)	SAM	QC
संख्यापूरणवाचक	(पूरण)	SAMY	QO
क्रियापद	(क्रिया)	several tags	VM
उपपद	(उपपद)		-
क्रियाविशेषण	(क्रिया_वि)	AVKV	RB
निषेधार्थक	(निषेध)	AVN	NEG
प्रश्नार्थक	(प्रश्न)	AVP	WQ
सम्बोधनम्	(सम्बो_सूचक)	-	INJ
उद्धरण	(उद्धरण)	-	UT
कृदन्त_क्रिया_सम्बन्ध	(कृ_क्रिया_सम्बन्ध)	AK	VM
साम्बन्धिक - सर्वनाम	(साम्ब_सर्व)	SNS	PRP
पदयोजक	(पदयोजक)	AVC	CC
वाक्ययोजक	(वाक्ययोजक)	-	CC
सादृश्यादि	(सादृश्यादि)	-	CC
द्विरुक्ति	(द्विरुक्ति)	-	RDP
भावसूचक	(भावसूचक)	UD	INJ
अन्य_भाषा	(अन्यभाषा)	AB	UNK
अज्ञात	(अज्ञात)	AB	UNK

Table 2: version 1.2

8 Modifications: Version 1.3

A review meeting to finalise the POS tags was held on 5th February, 2009. During the discussions it was emerged that it is necessary to provide examples for each of the tags, and also some guidelines for deciding the tags, in case of ambiguities. Further, two more tags – one for marking the punctuation marks and another for an **अव्यय** in case it does not fit in any of the given tags, were also added. The tagset emerged after the meeting is produced in Table 3.

9 POS tagging scheme for Sanskrit with Examples

संस्कृतवाक्यांशपदवर्गाङ्कनव्यवस्था[2]

क्रमाङ्कः	*वर्गः*	*सङ्केतः*	*उदाहरणम्*
1	नामसंज्ञा	नाम-सं	*Proper nouns such as names of human beings, rivers, place, books etc. - are to be marked as नाम-सं रामः/नाम-सं कृष्णः/नाम-सं च गोकुले/नाम-सं क्रीडतः। कृष्णः गोकुलात्/नाम-सं मथुराम्/नाम-सं गच्छति। कालिदासः/नाम-सं रघुवंशम्/नाम-सं कुमारसम्भवम्/नाम-सं च व्यरचयत्। रामायणम्/नाम-सं महाभारतम्/नाम-सं च इतिहासग्रन्थौ। सुमित्रानन्दः/नाम-सं गच्छति। *if सुमित्रानन्दः is a name of a person *If other words like देश, नगर etc are added to the proper noun will not be marked as नाम-सं। Eg: रामः अयोध्यानगरे वसति। *Here अयोध्या is a proper noun and not अयोध्यानगरम्/अयोध्यानगरी. So pos tag for अयोध्यानगरे is नाम
2	नामपदम्	नाम	*common nouns are to be marked as नाम गौः/नाम ग्रामम्/नाम स्वयम् याति। आचार्यः /नाम व्याकरणम्/नाम बोधयति। बालकः/नाम दाडिमफलम्/नाम खादति। सेवकः/नाम जलम्/नाम आनयति। इदम् पुस्तकम् पठनार्थम्/नाम स्वीकरोतु। गोविन्दः कृष्णानद्याम्/नाम स्नाति। सुमित्रानन्दः/नाम वनम् गच्छति। *if the word सुमित्रानन्दः refers to Lakshmana श्वेतः/नाम धावति। *Even though the word श्वेतः is a नामविशेषण, it is to be marked as नाम because of the absence of any other विशेष्यपदम्
3	नाम-विशेषणम्	नाम-वि	सुन्दरः/नाम-वि रामः जनककन्याम्/नाम-वि मनोरमाम्/नाम-वि सीताम् उवाह। वामनः/नाम-वि बालकः उन्नतम्/नाम-वि वृक्षम् आरोहति। तपोधनः/नाम-वि विष्णुभक्तः/नाम-वि नारदः ब्रह्मलोकम् गच्छति।

[2] 02 March 2009, Dept. of Sanskrit Studies, University of Hyederabad

			शूरः/नाम-वि सिंहः गहने/नाम-वि अरण्ये निवसति। प्रचण्डः/नाम-वि सूर्यः नीले/नाम-वि गगने प्रकाशते।
4	कृदन्त-नाम	कृ-नाम	रामः गतः/कृ-नाम सः माम् पृष्टवान्/कृ-नाम। सा विद्यालयम् गतवती/कृ-नाम। गच्छतः/कृ-नाम स्खलनम् क्वापि।
5	कृदन्तनामविशेषणम्	कृ-नाम-वि	गच्छन्तम्/कृ-नाम-वि पुरुषं निवारयेत्। गच्छन्/कृ-नाम-वि पिपीलिकः याति योजनानाम् शतानि। अगच्छन्/कृ-नाम-वि वैनतेयः अपि पदम् एकम् न गच्छति। भषन्तम्/कृ-नाम-वि शुनकम् मा ताडय। धावतः/कृ-नाम-वि अश्वात् पुरुषः अपतत्।
6	सर्वनाम	सर्व	सर्वे/सर्व सन्तु निरामयाः। अहं/सर्व संस्कृतमातरम् सेवे। तत्/सर्व सुन्दरम् नगरम्। त्वम्/सर्व मातरम् अनुगच्छ। कश्चित्/सर्व आगच्छति। *Even if कश्चित्, किञ्चित् etc. are avyayas, these are to be considered as सर्वनामाs, because these words are also referring some thing and the first part 'kim' can take different vibhaktis.
7	सर्वनाम-विशेषणम्	सर्व-वि	सर्वे/सर्व-वि जनाः सुखिनः भवन्तु। *A सर्वनाम which is a विशेषणम् is सर्वनामविशेषणम्। *Here the word सः modifies the noun, but it belongs to the category of सर्वनाम. So it is to be marked as सर्व-वि and not नाम-वि। सर्वे/सर्व-वि ऐक्यराज्यसमितिसदस्याः शान्तिप्रियाः। सः/सर्व-वि बालकः चतुरः। तत्/सर्व-वि पुस्तकम् कृष्णस्य भवति। कश्चित्/सर्व-वि बालकः आगच्छति।
8	साम्बन्धिक-सर्वनाम	साम्ब-सर्व	यद्/साम्ब-सर्व इह अस्ति तद्/साम्ब-सर्व अन्यत्र। *The सर्वनाम words which join two sentences. Eg: यत्-तत्। यस्य/साम्ब-सर्व न अस्ति स्वयम् प्रज्ञा शास्त्रम् तस्य/साम्ब-सर्व करोति किम्? यः/साम्ब-सर्व श्रमशीलः सः/साम्ब-सर्व कार्यसाधकः। ये/साम्ब-सर्व पञ्चतन्त्रम् पठन्ति ते/साम्ब-सर्व नीतिज्ञाः भवन्ति।
9	संख्यावाचकम्	संख्या	*Cardinal numbers बालकः:एकम्/संख्या द्वे/संख्यात्रीणि/संख्या इति वदति।
10	संख्येयवाचकम्	संख्येय	तिस्रः/संख्येय महिलाः कार्यम् कुर्वन्ति। द्वौ/संख्येय सैनिकौ युद्धम् कुरुतः।

			शतम्/संख्येय पुस्तकानि सन्ति। दशरथस्य तिस्रः/संख्येय भार्याः आसन्। आचार्यः त्रीणि/संख्येय पुस्तकानि दर्शयति। *Cardinal numbers which are विशेषणाs are संख्येयाs, else it is संख्या
11	संख्यापूरणवाचकम्	पूरण	कुन्तीपुत्रेषु प्रथमः/पूरण धर्मराजः। *(संख्याविशेषणाs) or Ordinal numbers/संख्यावाचकपदम् which denotes the order. शत्रुघ्नः दशरथस्य चतुर्थः/पूरण पुत्रः। एकादशे/पूरण उपवासः क्रियते। विंशतितमे/पूरण परीक्ड़ा अस्ति। रामस्य द्वितीयः/पूरण पुत्रः कुशः।
12	क्रियापदम्	क्रिया	रामः गच्छति/क्रिया वसति/क्रिया च वने। *Only तिङन्ताs are to be marked as क्रिया। गतवान्, पठितवान् etc. are treated as कृदन्त-नाम and not क्रिया। रामः हन्ति/क्रिया रावणम्। रामः अयोध्याम् याति/क्रिया। अग्निः दहति/क्रिया। विष्णुः गजेन्द्रम् मुमोच/क्रिया। पण्डितः अत्र बहुमन्यते/क्रिया। सः श्लाघ्यम् गमिष्यति/क्रिया। सः प्रसिद्धिम् याति/क्रिया।
13	निषेधार्थकम्	निषेध	आकाशः नीलः न/निषेध भवति। नो धत्ते जडताम् न/निषेध भगम् अयते। अलम्/निषेध चिन्तया। अलम्/निषेध विवादेन। मा/निषेध ते सङ्गः अस्तु अकर्मणि। संस्कृतम् कठिनम् न/निषेध भवति। त्वम् विवादम् मा/निषेध कुरु।
14	प्रश्नार्थकम्	प्रश्न	अपि/प्रश्न भवान् गच्छति? भवान् गच्छति किम्/प्रश्न? ननु/प्रश्न गच्छति भवान्? नारायणः सर्वान्तर्यामी किम्/प्रश्न?
15	उपपद-अव्ययम्	उपपद-अ	ग्रामम् परितः/उपपद-अ जलम् आवृतम्। *The अव्ययाs which decide the विभक्ति of its antecedent are the उपपद-अव्ययाs. ग्रामम् प्रति/उपपद-अ रथ्या प्रवर्तते। दुर्गम् अभितः/उपपद-अ परिखा अस्ति। कृष्णम् उभयतः/उपपद-अ गोपिकाः सन्ति। धर्मात् ऋते/उपपद-अ कुतः मोक्षः।
16	सम्बोधनसूचकम्	सम्बो-सूचक	हे/सम्बो-सूचक विधे क्रूरः असि त्वम्।

			*The indeclinables such as हे, भो etc. that are used to indicate सम्बोधन are सम्बोधनसूचक-अव्ययाs. भो/सम्बो-सूचक राम माम् उद्धर। *Here राम is not सम्बो-सूचक। अरे/सम्बो-सूचक बालक किम् करोषि ?
17	उद्धरणम्	उद्धरण	जननी जन्मभूमिः च स्वर्गात् अपि गरीयसी इति/उद्धरण उक्त्वा रामः विरराम। साप्तपदीनम् सख्यम् इति/उद्धरण ब्रह्मचारी पार्वतीम् अवोचत्। एवम्/उद्धरण मा करोतु इति/उद्धरण माता अतर्जयत्।
18	कृदन्तक्रियासम्बन्धी	कृ-क्रिया-सम्बन्धी	रमा फलम् कर्तयित्वा/कृ-क्रिया-सम्बन्धी अखादत्। छात्रः पठितुम्/कृ-क्रिया-सम्बन्धी गुरुकुलम् प्राविशत्। छात्रः पठन्/कृ-क्रिया-सम्बन्धी स्वपिति। भक्तः जलम् समर्प्य/कृ-क्रिया-सम्बन्धी सूर्यम् नमति। *कृदन्त words which modify the main verb are कृ-क्रिया-सम्बन्धीs. Eg: वसन् ददर्श। Here वसन् is कृ-क्रिया-सम्बन्धी।
19	क्रियाविशेषणम्	क्रिया-वि	मन्दम्/क्रिया-वि गच्छति। उच्चैः/क्रिया-वि भषति। सहसा/क्रिया-वि विदधीत न क्रियाम्। सः वेगेन/क्रिया-वि धावति। अश्वः शीघ्रम्/क्रिया-वि धावति। अश्वः वेगेन/क्रिया-वि धावति। सः सुखेन/क्रिया-वि वसति।
20	भावसूचकम्	भावसूचक	अहो/भावसूचक हास्यः खलु तपस्वी ? कर्णः। हा/भावसूचक हन्त/भावसूचक लुण्ठाकेन सर्वम् अपहृतम्।
21	पद-योजकम्	पदयोजक	रामः भीमः च/पदयोजक वीरतमौ। *The indeclinable (अव्ययम्) which joins two or more words भवान् गच्छति पठति वा/पदयोजक— ज्ञाने धर्मः उत/पदयोजक प्रयोगे।
22	द्विरुक्तिः	द्विरुक्ति	वधूः मन्दम्/क्रिया-वि मन्दम्/द्विरुक्ति चलति। गजः शनैः/क्रिया-वि शनैः/द्विरुक्ति गच्छति। *In द्विरुक्ति first part is given its own category and second word is marked as द्विरुक्ति। हा/भावसूचक हा/द्विरुक्ति देवि स्फुटति हृदयम्। क्रुद्धः रामः/नाम-सम् रामः/द्विरुक्ति इति वदति। *वसन्तकाले सम्प्राप्ते काकः काकः पिकः पिकः। In this sentence काकः काकः or पिकः पिकः are not द्विरुक्ति। All four of them are नामपदs.
23	सादृश्यम्	सादृश्य	चन्द्र इव/सादृश्य मुखम् अस्ति। यथा/सादृश्य राजा तथा/सादृश्य प्रजा।

24	अवधारणम् (निश्चयात्मकम्)	अवधारण	रामः एव/अवधारण प्रियदर्शनः।
25	वाक्य-योजकम्	वाक्य-योजक	केशवः गृहं गच्छति रामः च/वाक्य-योजक पठति। *The indeclinables which connect two verbs. अध्यापकः पाठयति छात्राः पठन्ति च/वाक्य-योजक— ते गतवन्तः रामः अपि/वाक्य-योजक गच्छेत्। यत्र/वाक्य-योजक योगेश्वरः कृष्णः तत्र/वाक्य-योजक पार्थः धनुर्धरः। यद्यपि/वाक्य-योजक तेन परिश्रमः कृतः तथापि/वाक्य-योजक सफलः न जातः। यथा/वाक्य-योजक आज्ञापयति भवान् तथा/वाक्य-योजक करोमि। यदि/वाक्य-योजक मेघः वर्षति तर्हि/वाक्य-योजक मयूरः नृत्यति।
26	अव्ययम्	अव्यय	अद्य/अव्यय सः गृहम् गच्छति। *If an अव्ययम् does not fit in any of the above categories then it is marked as अव्यय। वायुः सर्वत्र/अव्यय वर्तते। कः नु/अव्यय अस्मिन् साम्प्रतम् लोके।
27	उपाधिः	उपाधि	Dr./उपाधि— विद्वान्/उपाधि रामानन्दः। *The words which indicate designation such as Dr, sri etc.
28	अन्य-भाषा	अन्य-भाषा	JOHN/अन्यभाषा *To mark the words from other languages
29	चिह्नम्	चिह्न	भवान् किम् करोति ।/चिह्न हे राम !/चिह्न माम् उद्धर। *Punctuation marks only
30	संशयः	?	*In case of doubt put a ‘?’ #When you have a doubt as to whether the tag is A or B, put A/B.
31	अज्ञातम्	अज्ञात	Words which do not fall under any of these categories.

Table 3: Ver 1.3 with examples

#It was found that the human annotators have sometimes confusion between a couple of tags. Marking them as just UNK (अज्ञात) is not a good solution. Because here the annotators know for sure that the word belongs to one of the two/three tags. But he is confused. In such cases we have provided this facility, which later may be resolved during discussions. This tag thus will not occur in any Gold standard annotated data.

10 Comparison with the ILMT Tagset

The table 4 below gives a list of Sanskrit POS tags and the corresponding ILMT tags.

10.1 Tags not used from ILMT

Following tags from ILMT tagset are not used in Sanskrit as they are not needed.

- NLoc
- Verb Aux
- Post Position
- Particles
- Classifier
- Compounds
- Echo

10.2 New tags necessary for Sanskrit

Following extra tags are needed for Sanskrit, in addition to the tags from ILMT.

- उपसर्ग (upasarga)
  This tag is needed only to handle Vedic texts where upasargas are written separately.
- उपपद अव्ययम् (upapada avyayam)
  This tag is needed to tag the indeclinables which govern the vibhakti of the noun to which it is attached. For example, 'saha' will assign 3^{rd} vibhakti to the noun as in 'रामेण सह'.
  (This upapada vibhakti is subset of the 'PSP' in ILMT tagset.)
- अव्ययम् (avyayam)
  In case a word is an avyaya, and can not be marked by any of the tags in the tagset, it is marked as an avayaya.

10.3 One-Many mappings between ILMT tagset and Sanskrit tagset

Some of the ILMT tagsets were ambiguous/coarse-grained from Sanskrit analysis point of view, and hence these tags were sub-classified further. Table 5 gives the ILMT tags which are split and the corresponding Sanskrit sub-classification.

sr no	*Description*	*SKT ABBR*	*ILMT Tag*	Eng ABBR
1	नामपदम्	(नाम)	NN	NP
2	नामसंज्ञा	(नाम_सं)	NNP	NS
3	नाम-विशेषणम्	(नाम_वि)	JJ	NV
4	कृदन्तनाम	(कृ_नाम)	NN	KN
5	कृदन्तनाम-विशेषणम्	(कृ_नाम-वि)	JJ	KNV
6	सर्वनाम	(सर्व)	PRP	SN
7	सर्वनाम - विशेषणम्	(सर्व_वि)	DEM	SNV
8	साम्बन्धिक - सर्वनाम	(साम्ब_सर्व)	PRP	SSN
9	संख्यावाचकम्	(संख्या)	QC	SMK
10	संख्येयवाचकम्	(संख्येय)	QC	SMKY
11	संख्यापूरणवाचकम्	(पूरण)	QO	SMKP
12	उपसर्ग	(उप)(Only in Vedic Sanskrit)	-	-
13	क्रियापदम्	(क्रिया)	VM	KP
14	निषेधार्थकम्	(निषेध)	NEG	AN
15	प्रश्नार्थकम्	(प्रश्न)	WQ	AP
16	उपपद अव्ययम्	(उपपद-अ)	-	AUP
17	सम्बोधनम् सूचकम्	(सम्बो_सूचक)	INJ	ASMB
18	उद्धरणम्	(उद्धरण)	UT	AUD
19	कृदन्त_क्रिया_सम्बन्धी	(कृ_क्रिया_सम्बन्धी)	VM	KKS
20	क्रियाविशेषणम्	(क्रिया_वि)	RB	KV
21	भावसूचकम्	(भावसूचक)	INJ	ABS
22	पदयोजकम्	(पदयोजक)	CC	APY
23	द्विरुक्तिः	(द्विरुक्ति)	RDP	DV
24	सादृश्यादि	(सादृश्य)	CC	ASD
25	निश्चयात्मकम्(अवधारण)	(अवधारणा)	INTF	INTF
26	वाक्ययोजकम्	(वाक्य-योजक)	CC	AVK
27	अव्ययम्	(अव्यय)	-	A
28	उपाधिः	(उपाधि)	RP?	U
29	अन्य_भाषा	(अन्य-भाषा)	UNK	AB
30	चिह्नम्	(चिह्न)	SYM	PUNC
31	अज्ञात	(अज्ञात)	UNK	UNK
32	*संशयः	(?)	-	(?)

Table 4: version 1.3 dated 2nd March 2009

ILMT tag	*SKT tags*
JJ	ना वि and कृ-ना-वि
NN	नाम and कृ नाम
PPR	सर्व and साम्ब-सर्व
QC	संख्या and संख्येय
INJ	सम्बोधन and भावसूचक
CC	पदयोजक, वाक्य-योजक and सादृश्य
UNK	अन्य-भाषा and अज्ञात

Table 5: ILMT POS tags subclassified for Sanskrit

11 Members of the Consortium

Following are the members of the Sanskrit Consortium

- Amba Kulkarni, University of Hyderabad
- Girish Nath Jha, JNU, Delhi
- V N Pandurangi, JRRSU, Jaipur
- Tirumala Kulkarni, Poornaprajna Vidyapeetha, Bangalore
- S S Murty, RSVP, Tirupati
- Shrinivasa Varakhedi, Sanskrit Academy, Hyderabad
- Dipti Misra Sharma, IIIT, Hyderabad

Special Invitees:

- Prof. K N Murty, Univ of Hyderabad
- Shri. Kumar Swami, Retd. Principal, KV

Project Co-investigators and Linguists:

- Dr. Varalakshmi, Sanskrit Academy, Hyderabad
- Dr. Rajdhar Mishra, JRRSU, Jaipur
- Dr. V Sheeba, Univ of Hyderabad
- Dr. Devanand Shukl, Univ of Hyderabad
- Dr. R Chandrashekhar, JNU
- Dr. Ramchandra, PV, Bangalore