Inflectional morphology analyser for Sanskrit

Girish Nath Jha, Muktanand Agrawal, Sudhir K. Mishra, Diwakar Mani, Diwakar Mishra, Manji Bhadra, Surjit K. Singh, - Subash

Inflectional Morphology Analyzer for Sanskrit

Girish Nath Jha, Muktanand Agrawal, Subash, Sudhir K. Mishra, Diwakar Mani, Diwakar Mishra, Manji Bhadra, Surjit K. Singh

girishj@mail.jnu.ac.in

Special Centre for Sanskrit Studies
Jawaharlal Nehru University
New Delhi-110067

The paper describes a Sanskrit morphological analyzer that identifies and analyzes inflected noun-forms and verb-forms in any given sandhi-free text. The system which has been developed as java servlet RDBMS can be tested at http://sanskrit.jnu.ac.in (Language Processing Tools > Sanskrit Tinanta Analyzer/Subanta Analyzer) with Sanskrit data as unicode text. Subsequently, the separate systems of subanta and tiṅanta will be combined into a single system of sentence analysis with karaka interpretation. Currently, the system checks and labels each word as three basic POS categories - subanta, tiṅanta, and avyaya. Thereafter, each subanta is sent for subanta processing based on an example database and a rule database. The verbs are examined based on a database of verb roots and forms as well by reverse morphology based on Pāṇinian techniques. Future enhancements include plugging in the amarakosha (http://sanskrit.jnu.ac.in/amara) and other noun lexicons with the subanta system. The tiṅanta will be enhanced by the kṛdanta analysis module being developed separately.

1. Introduction

The authors in the present paper are describing the *subanta* and *tiṅanta* analysis systems for Sanskrit which are currently running at http://sanskrit.jnu.ac.in. Sanskrit is a heavily inflected language, and depends on nominal and verbal inflections for communication of meaning. A fully inflected unit is called *pada*. The *subanta padas* are the inflected nouns and the *tiṅanta padas* are the inflected verbs. Hence identifying and analyzing these inflections are critical to any further processing of Sanskrit.

The results from the *subanta* analyzer for the input text fragment

आम्रस्य आत्मकथा
चपलाः बालकाः आम्राणाम् उद्यानं गच्छन्ति । तत्र आम्रफलानि पश्यन्ति प्रसन्नाः च भवन्ति ।

are displayed as follows –

आम्रस्य_SUBANTA_SUBANTA आत्मकथा [आत्मकथा (स्त्रीलिङ्ग) + सु, प्रथमा, एकवचन] [*_PUNCT] चपलाः [चपल (पुल्लिङ्ग) + जस्, प्रथमा, बहुवचन] बालकाः [बालक (पुल्लिङ्ग) + जस्, प्रथमा, बहुवचन] आम्राणाम् [आम्र (पुल्लिङ्ग) + आम्, षष्ठी, बहुवचन] उद्यानं [उद्यन्+अम्, द्वितीया, एकवचन] [गच्छन्ति_VERB] [।_PUNCT] [तत्र_AV] आम्रफलानि [आम्रफल+जस्/शस् प्रथमा/द्वितीया, बहुवचन] [पश्यन्ति_VERB] प्रसन्नाः [प्रसन्न (पुल्लिङ्ग) + जस्, प्रथमा, बहुवचन] [च_AV] [भवन्ति_VERB] [।_PUNCT]

Those colored blue are non *subanta* categories and those colored red are possible errors. The default black colored ones are the *subanta* padas analyzed.

The word गच्छन्ति from the above input text resulted in the following output from the *tiṅanta* analyzer system –

गच्छन्ति { (कर्तृवाच्य) गम् ([भ्वादिगण] [अनिट्] [सकर्मक]) ([लट्]) झि ([परस्मै] [प्रथम–पुरुष] [बहुवचन]) }

2. The *Subanta* Anlyzer

The system accepts unicode (UTF-8) sandhi-free Devanagari Sanskrit inputs (word, sentence or text) and processes it according to the following sequence -

INPUT TEXT
↓
PRE-PROCESSOR
↓
VERB DATABASE→ LIGHT POS TAGGING ←AVYAYA DATABASE
↓
SUBANTA RECOGNIZER←VIBHAKTI DATABASE
↓
SUBANTA RULES→ *SUBANTA ANALYZER* ←SANDHI RULES
↓
SUBANTA ANALYSIS

The PREPROCESSOR does the simplification and normalization of the Sanskrit text (for example, deletes Roman characters, other invalid words, punctuations etc). The POS TAGGER identifies four categories **AV**yaya, **VERB**, **PUNCT**uation and **SUBANTA**. The SUBANTA RECOGNIZER does *vibhakti* identification and isolation by searching the *vibhakti* database. The SUBANTA ANALYZER does analysis by checking the *subanta* rule base and sandhi rules. Analysis includes splitting the NPs into its constituents - base [(*prātipadika*) (PDK)], case-number markers (*kāraka-vacana-vibhakti*).

3. Sanskrit sentence and basic POS categories

A Sanskrit sentence has NPs (including AVs), and VPs. Cardona[1] (1988) defines a sentence as -

$$(N - E^{n})\ p\ \ldots\ (V - E^{v})\ p$$

After *sup* and *tiṅ* combine with PDK, they are assigned syntactico-semantic relation by the *kāraka* stipulations to return complete sentences.

[1] George Cardona, 1988 Pāṇini, His Work and its Traditions, vol **...** i (Delhi: MLBD, 1988)

3.1 Sanskrit *subanta* (inflected nouns)

Sanskrit nouns are inflected with seven case markers in three numbers. Potentially, a noun can be declined in all three genders. Sanskrit noun forms can be further complicated by being a derived noun as primary (*kṛdanta*), secondary (*taddhitānta*), feminine forms (*strīpratyayānta)* and compounds (*samāsa).* They can also include *upasargas* and AVs etc. According to Pāṇini, there are 21 case suffixes called *sup* (seven *vibhaktis* combined with three numbers)[2], which can attach to the nominal bases (PDK) according to the syntactic category, gender and end-character of the base. Pāṇini has listed these as sets of three as:

su, au, jas
am, auṭ, śas
ṭā, bhyām, bhis
ṅe, bhyām, bhyas
ṅasi, bhyām, bhyas
ṅas, os, ām
ṅi, os, sup[3]

for singular, dual and plural[4] respectively. These suffixes are added to the PDKs[5] (any meaningful form of a word, which is neither a root nor a suffix) to obtain inflected forms NPs. PDKs are of two types: primitive and derived. The primitive bases are stored in g*aṇapāṭha* [(GP) (collection of bases with similar forms)] while the latter are formed by adding the derivational suffixes. NPs are of mainly six types –

3.1.1 *avyaya subanta* (indeclinable nouns)

Avyaya subanta, remain unchanged under all morphological conditions[6]. According to Pāṇini [2.2.82][7], affixes *cāp, ṭāp, ḍāp,* (feminine suffixes) and *sup* are deleted by *luk* when they occur after an AVs. Pāṇini defines AVs as *svarādinipātamavyayam* [1.1.36]*, kṛnmejantaḥ* [1.1.38]*, ktvā tosun kasunaḥ* [1.139] and *avyayībhāvaśca* [1.1.40][8] etc.

3.1.2 basic *subantas* (primitive nouns)

Basic *subantas* are formed by primitive PDKs found in the Pāṇini's *gaṇapāṭha*. For our purpose, all those nouns, the base or inflected form of which can be found in a lexicon can be considered basic *subantas*. Sometimes, commonly occurring primary or secondary derived nouns, feminine or compound forms can also be found in the lexicon. Therefore such *subantas* are also considered basic and do not require any reverse derivational analysis unless specifically required.

---

[2] स्वौजसमौट्छष्टाभ्याम्भिस्ङेभ्याम्भ्यस्ङसिभ्याम्भ्यस्ङसोसांङ्योस्सुप्

[3] सुपः

[4] द्व्येकयोर्दिवचनैकवचने

[5] अर्थवधातुरप्रत्ययः प्रातिपदिकम् ।१।२।४५।, कृत्तद्धितसमासाश्च ।१।२।४६।।

[6] सदृशं त्रिषु लिङ्गेषु सर्वासु च विभक्तिषु ।
वचनेषु च सर्वेषु यन्न व्येति तदव्ययम् ।। [गोपथ ब्राह्मण]

[7] अव्ययादाप्सुपः [२.४.८२]

[8] स्वरादिनिपातमव्ययम् [१.१.३६], कृन्मेजन्तः [१.१.३८], क्त्वा–तोसुन्–कसुनः [१.१.३९], अव्ययीभावश्च [१.१.४०]

Such inflected nouns are formed by inflecting the base or PDKs (*arthavadadhāturapratyayaḥ prātipadiakam)* with *sup*. For example: *rāmaḥ, śyāmaḥ, pustakālayaḥ, vidyālayaḥ* etc.

3.1.3 *samāsānta subanta* (compound nouns)

Simple words (*padas*), whether substantives, adjectives, verbs or indeclinables, when added with other nouns, form *samāsa* (compound). Sanskrit *samāsas* are divided into four categories, some of which are divided into sub-categories. The four main categories of compounds are as follows:

- adverbial or *avyayībhāva*,
- determinative or *tatpuruṣa,*
- attributive or *bahuvrīhi* and
- copulative or *dvandva*. *dvandva* and *tatpuruṣa* compounds may be further divided into sub-categories

3.1.4 *kṛdanta subanta* (primary derived nouns)

The primary affixes called *kṛt* are added to verbs to derive substantives, adjectives or indeclinables.

3.1.5 *taddhitānta subanta* (secondary derived nouns)

The secondary derivative affixes called *taddhita* derive secondary nouns from primary nouns. For example - *dāśarathī, gauṇa* etc.

3.1.6 *strīpratyayānta subanta* (feminine derived nouns)

Sanskrit has eight feminine suffixes *ṭāp, cāp ḍāp, ṅīṣ, ṅīn, ṅīp, uṅ* and *ti* etc. and the words ending in these suffixes are called *strīpratyayānta* For example - *ajā, gaurī, mūṣikā, indrāṇī, gopī, aṣṭādhyāyī, kurucarī, yuvatī, karabhorū* etc.

4. Recognition of Sanskrit *subanta*

4.1 Recognition of punctuations

System recognizes punctuations and tags them with the label PUNCT. If the input has any extraneous characters, then the input word will be cleaned from these elements (i.e. 'normalized') so that only Devanāgarī Sanskrit input text is sent to the analyzer. For example, "रा/&^%@#मः, बा,'":–=लकः" ➔ रामः, बालकः

4.2 Recognition of *Avyayas*

System takes the help of *avyaya* database for recognizing AVs. If an input word is found in the AVs database, it is labeled AV, and excluded from the *subanta* analysis as AVs do not change forms after *subanta* affixation. We have stored most AVs in the *avyaya* database.

4.3 Recognition of verbs

System takes the help of verb database for verb recognition. If an input is found in the verb database, it is labeled VERB and thus excluded from *subanta* analysis. Since storing all Sanskrit verb forms is not possible, we have stored verb forms of commonly used 450 verb roots.

4.4 Recognition of *subanta*

Thus, a process of exclusion identifies the nouns in a Sanskrit text. After the punctuations, *avyayas* and verbs are identified, the remaining words in the text are labeled SUBANTA.

5. Analysis of *subanta*

System does analysis of inflected nouns with the help of two relational database - examples and rules. Brief description of these databases follows-

5.1 Example database

All complicated forms (which are not analyzed according to any rule) including those of some pronoun are stored the database. For example: अहम्=अस्मद+सु प्रथमा एकवचन;अहं=अस्मद+सु प्रथमा एकवचन;आवाम्=अस्मद+औ प्रथमा द्वितीया द्विवचन;आवां=अस्मद+औ प्रथमा द्वितीया द्विवचन;वयम्=अस्मद+जस प्रथमा बहुवचन;वयं=अस्मद+जस प्रथमा बहुवचन;माम्=अस्मद+अम द्वितीया एकवचन;मां=अस्मद+अम द्वितीया एकवचन

5.2 Rule database

The *subanta* patterns are stored in this database. This database analyzes those nouns which match a particular pattern from the rule base. For example, रामः, नदी, रमा, पुस्तकम etc. First, the system recognizes *vibhakti* as the end character of nouns. For example, '**ः**' is found in nominative singular (1-1) like -रामः, श्यामः, सर्वः, भरतः एकः . The system isolates '**ः**' and searches for analysis in the *sup* rule base. In the case of nominative and accusative dual (1-2/2-2), PDK forms will be 'ौ' ending, for example - रामौ, श्यामौ, सर्वौ, एकौ. The system isolates 'ौ' and searches for analysis by matching in the rule database. The sample data is as follows –

ा=ा+सु प्रथमा एकवचन;ाभ्याम्=+भ्याम् तृतीया चतुर्थी पञ्चमी द्विवचन;ाभ्यां=+भ्याम् तृतीया चतुर्थी पञ्चमी द्विवचन;भ्याम्=+भ्याम् तृतीया चतुर्थी पञ्चमी द्विवचन;भ्यां=+भ्याम् तृतीया चतुर्थी पञ्चमी द्विवचन;ेभ्यः=+भ्यस् चतुर्थी पञ्चमी बहुवचन;भ्यः=+भ्यस् चतुर्थी पञ्चमी बहुवचन;

5.3 verb data sample

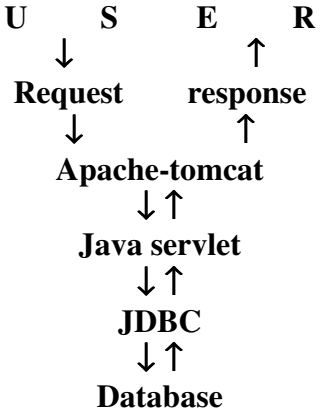
भवति,भवतः,भवन्ति,भवसि,भवथः,भवथ,भवामि,भवावः,भवामः,भवतु,भवताम्,भवन्तु,भव,भवतम्,भवत,भवानि,भवाव,भवाम,अभवत्,अभवताम्,अभवन्,अभवः,अभवतम्,अभवत,अभवम्,अभवाव,अभवाम,भवेत्,भवेताम्,भवेयुः,भवेः,भवेतम्,भवेत,भवेयम्,भवेव,भवेम,बभूव,बभूवतुः,बभूवुः,बभूविथ,बभूवथुः,बभूव,बभूव,बभूविव

5.4 *avyaya* data sample

अ,कश्चित्,सदैव,अकस्मात्,अकाण्डे,अग्निसात्,अग्नी,अघोः,अङ्ग,अजस्रम्,अञ्ज्सा,अतः,अति,अतीव,अत्र,अथ,अथकिम्,अथवा,अथो,अद्धा,अद्य,अद्यापि,अधरात्,अधरेद्युः,अधरेण,अधः,अधस्तात्,अधि,अधिहरि,अधुना,अधोऽधः,अध्यर्धानतः,अनिशम्,अनु,अनेकधा,अनेकशः,अन्तः,अन्तरा,अन्तरेण,अन्यतः,अन्यत्,अन्यत्र

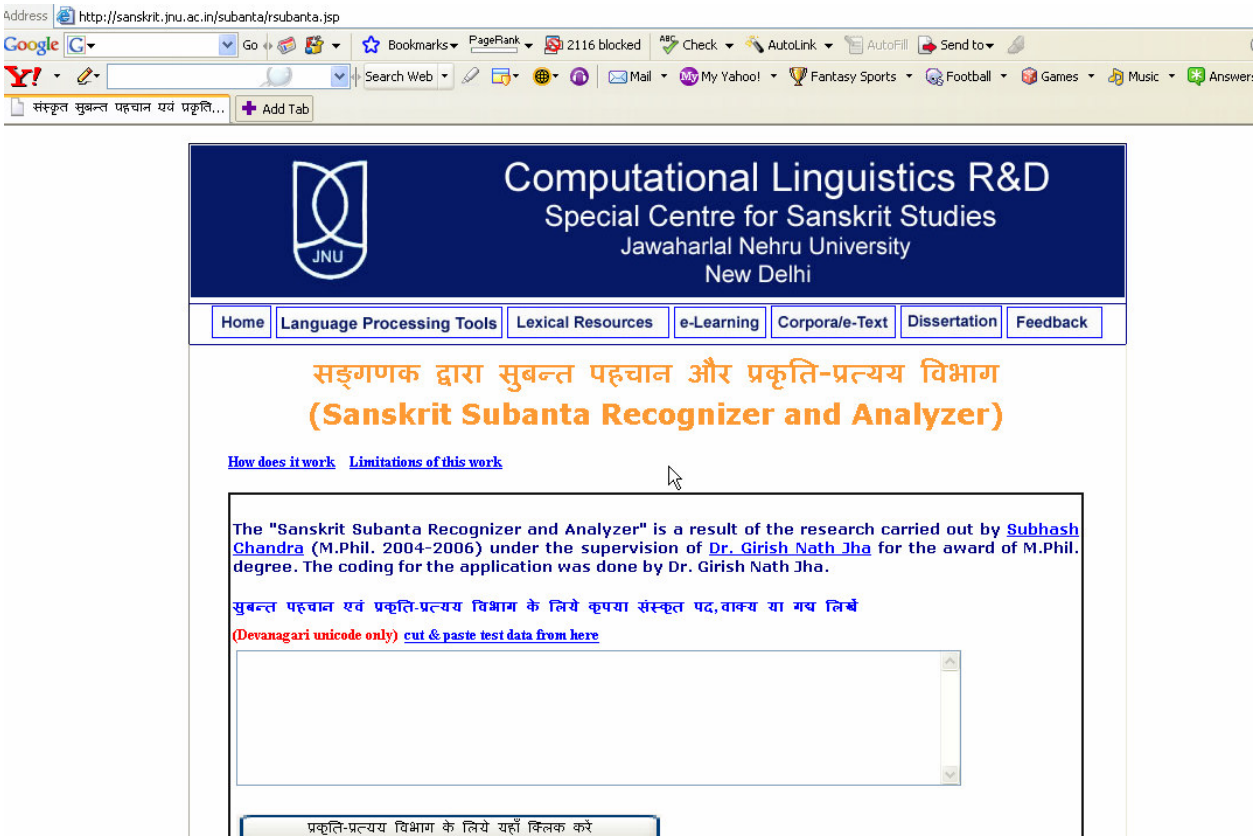
5.5 Architecture of the system

The following model describes the interaction between multi-tiered architecture of the *subanta* analyzer:



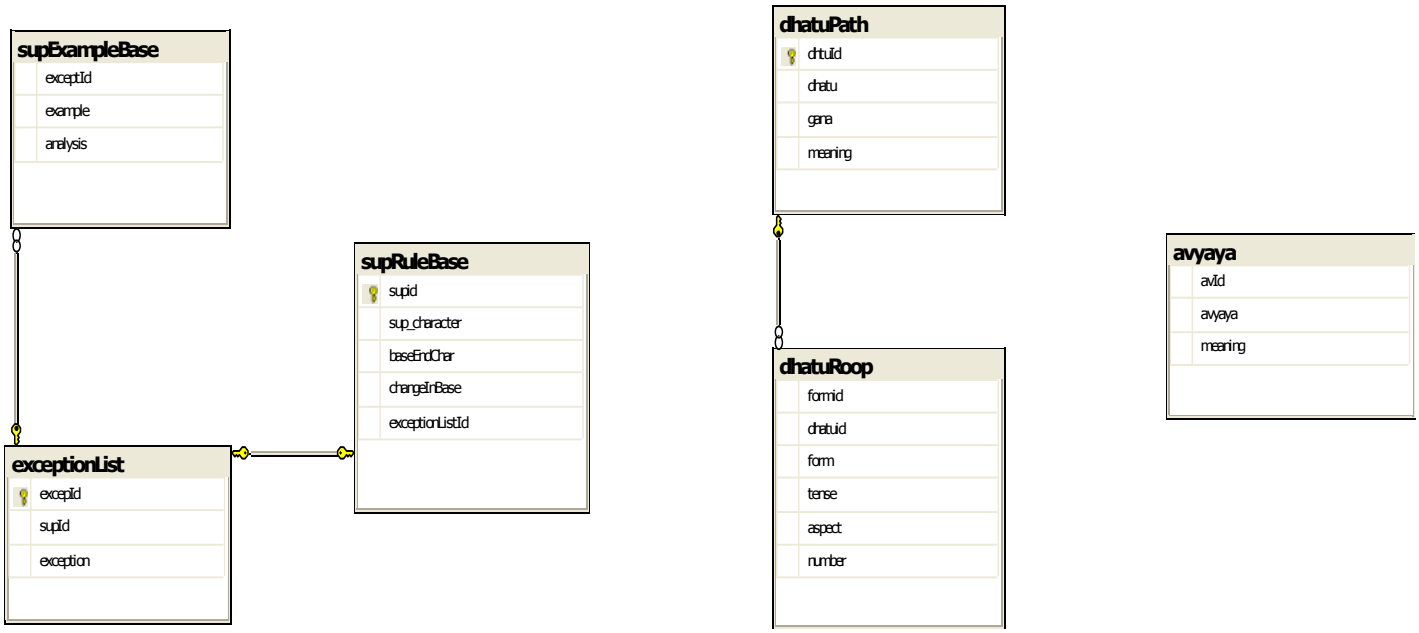
5.6 Front-end: online interface

The Graphical User Interface (GUI) is produced by JSP (Java Server Pages). The JSP interface allows the user to give input in Devanagari utf-8 format using HTML text area component. The user interface is displayed as follows:



5.7 Back-end: database / txt files

There are two versions of the system; the server-based version connects to a MS-SQL Server 2005 RDBMS through JDBC. The rule base, example base and other linguistic resources are stored as Devanagari utf-8. The PC based portable version, for obvious reasons, cannot have RDBMS support. Therefore, we have our rules and data stored in utf-8 text files as backend. A design of the reverse *subanta* database is given below-



The *supRuleBase* table has relations with the *exceptionList* table. Any exception figuring in the rule base must have a description in the exception list. The table *supExampleBase* depends on the *exceptionList* and must provide analysis for each example figuring in the *exceptionList* and marked in the *supRuleBase.* The *dhāturūpa* object depends on the *dhātupāṭha* object while the AVs is a floating object as of now. These linguistic resources are checked for recognition of nouns, and the rules and example bases are searched for analysis. System uses some text/data files whose samples have been given in earlier sections.

5.2 Database connectivity

The database connectivity is done through Java Database connectivity (JDBC) driver. JDBC Application Programming Interface (API) is the industry standard for database independent connectivity for Java and a wide range of SQL databases. JDBC technology allows using the Java programming language to develop 'write once, run anywhere' capabilities for applications that require access to large-scale data. JDBC works as bridge between Java program and Database. SQL server 2005 and JDBC support input and output in Unicode, so this system accepts unicode Devanagari text as well as prints result in unicode Devanagri too[9].

6. Limitations of the system

6.1 Limitations of the recognition process

This system has the following recognition limitations:

- at present, we have approximately verb forms for only 450 commonly found verb roots in the verb database. Though it is very unlikely that ordinary Sanskrit literature will overshoot this list, yet the system is likely to start processing verb forms as nouns if not found in this limited database.
- at this point, the system will wrongly mark prefixed or derived verb-forms as nouns as they will not be found in the verb database. The gains from the *tiṅanta* analyzer will be added here shortly to overcome this limitation.
- currently this work assumes sandhi-free text. So, a noun or verb with sandhi is likely to return wrong results. The gains from a separate research on sandhi processing will be used to minimize such errors.
- currently, our AV database has only 519 AVs. It is not enough for AV recognition in ordinary Sanskrit literature. In this case, the system is likely to start processing AVs as nouns, if it is not found in AVs database.
- some forms ending in primary affixes look like nouns while they are AVs. For example: पठितुम्, गत्त्वा, आदाय, विहस्य etc. System will incorrectly recognize and process them as *subantas*.
- many nouns (for example, *śatṛ pratyayānta* in locative singular) look like verbs. These will be wrongly recognized as verbs for example: भवति, गच्छति, पठति, चलति etc. To solve this problem, we will have a hybrid POS category called SUPTIN for those verb forms which are *subantas* as well.

[9] http://java.sun.com/products/servlet/

6.2 Limitations of the analysis process

The system has the following analysis limitations:

- same forms are available in the dual of nominative and accusative cases, for example, रामौ, dual of instrumental, dative and ablative cases, for example रामाभ्याम्, plural of dative and ablative cases, for example रामेभ्यः, dual of genitive and locative cases, for example रामयोः. In neuter gender as well, the nominative and accusative singular forms may be identical as in पुस्तकम् (1-1 and 2-1). In such cases, the system will give all possible results as in

रामौ	=	औ	[प्र./ द्वि. द्विव.]
रामाभ्याम्	=	भ्याम्	[तृ./च./पं. द्विव.]
रामेभ्यः	=	भ्यस्	[च./पं. बहुव.]
रामयोः	=	ओस्	[ष./स. द्विव.]
पुस्तकम्	=	सु/अम्	[प्र./द्वि. एकव.]
हरेः	=	ङसि/ङस्	[पं./ष. एकव.]

- some *kṛdanta* forms (generally *lyap, tumun,* and *ktvā* suffix ending) look like nouns (for example - विहस्य पठित्वा, गत्वा, पठितुम्, गन्तुम, नेतुम्, प्रदाय, विहाय etc.). In such cases, the system may give wrong results as:

विहस्य = विह + ङस् षष्ठी एकवचन
पठित्वा = पठित्वा + सु प्रथमा एकवचन
गत्वा = गत्वा + सु प्रथमा एकवचन
पठितुम् = पठितु + अम् द्वितीया एकवचन
गन्तुम् = गन्तु + अम् द्वितीया एकवचन

- at this point, system does not have gender information for all PDKs, nor does it attempt to guess the gender. This limitation is going to be minimized by plugging in the *amarakośa* shortly.

- currently this system is giving multiple results in ambiguous cases, because the words as analyzed a single tokens. This will be solved by adding the gains from the research on *kāraka* and gender of nouns which concluded recently.

7. The *tiṅanta* analyzer

Verbs constitute an important part of any language. A sentence indispensably requires a verb to convey complete sense. Given the importance of verb and verb phrases in any linguistic data, it is necessary to develop a proper strategy to analyze them. Creating lexical resource for verbs along with other parts of speech is a necessary requirement. Sanskrit is a highly inflectional language. It is relatively free word-order language. The semantic inter-relation among the various components of a sentence is established through the inflectional suffixes.

Scholars have done efforts to analyze Sanskrit verb morphology, both in theory and in computation. Some of the major works are listed below:

- Gerard Huet has developed a lemmatizer that attempts to tag inflected Sanskrit verbs along with other words. This lemmatizer knows about inflected forms of derived stems which are not apparent in the display of the main stem inflection. It, however, does not attempt to lemmatize verbal forms with pre-verbs but only invert root forms. The site also provides a long list of the conjugated forms of verb-roots in the present, imperfect, imperative, optative, perfect, aorist and future tenses as a PDF document.
- *Prajna* project of ASR Melkote claims to do module generation and analysis of 400 important Sanskrit roots in three voices (Active, Passive and Impersonal), 10 lakāra, 6 tense and 4 moods,
- Aiba (2004) claims to have developed a Verb Analyzer for classical Sanskrit which can parse Sanskrit verb in Present, Aorist, Perfect, Future, Passive and Causative forms. This site actually works only for some verbs and accepts that the results are not reliable,
- *Desika* project of TDIL, Govt. of India claims to be an NLU system for generation and analysis for plain and accented written Sanskrit texts based on grammar rules of Pāṇini's *Aṣṭādhyāyī*. It also claims to have a database based on *Amarakośa* and heuristics based on *Nyāya* & *Mīmāmsā Śāstras* and claims to analyze Vedic texts as well,
- RCILTS project at SC&SS, JNU has reportedly stored all verb forms of Sanskrit in a database,
- *Śābdabodha* project of TDIL, govt. of India claims to be an interactive application to analyze the semantic and syntactic structure of Sanskrit sentences,
- The ASR Melcote website reports that a Sanskrit Authoring System is under development at C-DAC Bangalore. The system is supposed to make making tools for morphological, syntactic and semantic analyses with word split programs for sandhi and *samāsa*.
- Cardona (2004) discussed Pāṇini's derivational system involving aspect of linguistics, grammar and computer science.
- Whitney (2002) listed all the quotable roots of the Sanskrit language together with the tense and the conjugation system.
- Mishra and Jha (2004) describe a module (Sanskrit *Kāraka* Analyzer) for identification and description of *kāraka* according to *Pāṇinian kāraka* formulations.
- Edgren (1885) discussed verb roots of Sanskrit language according to Sanskrit grammarians.
- Joshi (1962) presented linguistic analysis of verb and nouns of Sanskrit language

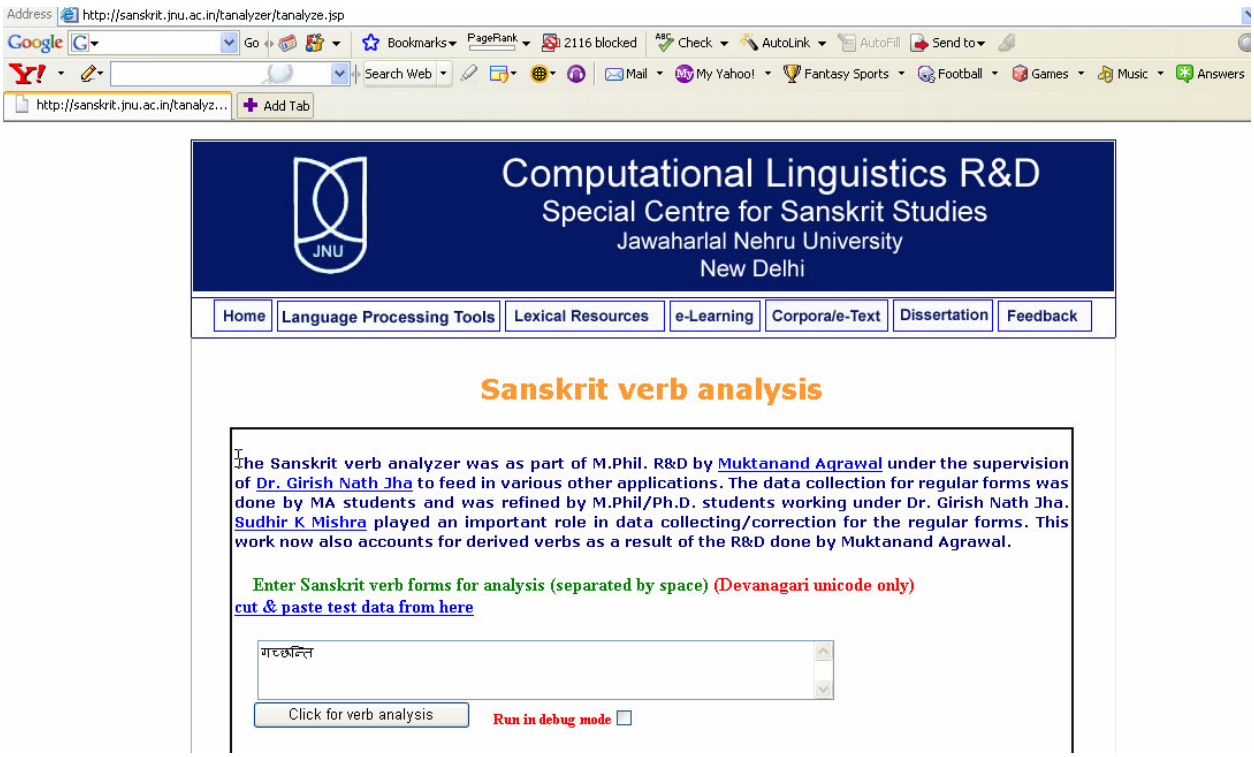
- Jha and Mishra (2004) proposed *a model for Sanskrit verb inflection identification that would correctly describe verbs in a laukika Sanskrit text.* They presented a module to identify the verb by applying Pāṇini rules in reverse with the help of a relational database.

This module can also be used to identify the types of sentences as active or passive voice with complete reference of the verb.

Present work, which owes a lot to above listed efforts, has some specific features such as:

- The system takes into account the *Pāṅinian* analysis and develops its methodology by applying it into reverse direction.
- It aims at developing a comprehensive strategy so that any *tiṅanta* can be analyzed with the same technique.
- It can be further expanded and modified to recognition and analysis of denominatives
- it is an online servlet-unicode database system with input-output in unicode only

The front-end of the *tiṅanta* analyzer is as follows -

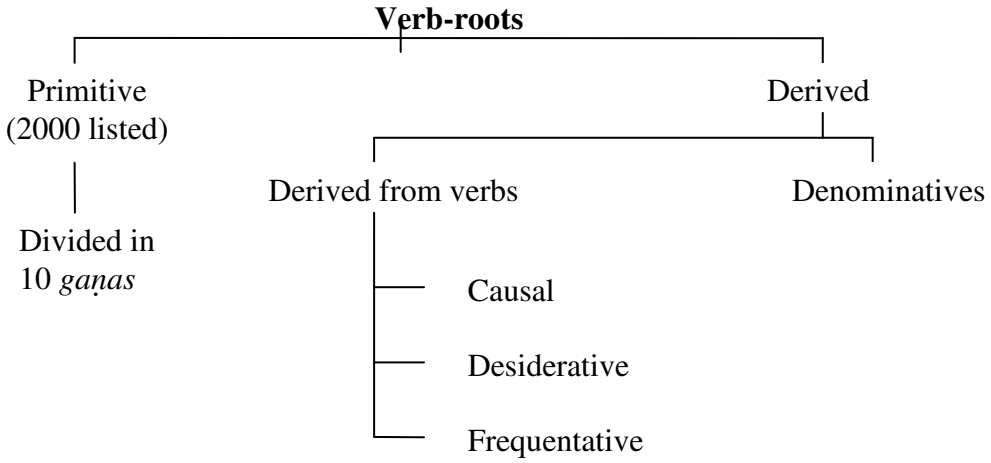


8. Sanskrit verb-morphology

Verbs have been of central importance to Sanskrit grammarians. *Yāska* insisted so much on them that he propounded that all the nominal words are derived from verb roots. Verbs convey the sense of becoming[10]. Sanskrit follows a well defined process in the formation of *padas*. Both noun *padas (subanta)* as well as verb *padas (tiṅanta)* have to undergo certain inflectional

[10] *bhāvapradhānamākhyātam (Yāska, Nirukta)*

processes in which various nominal or verbal affixes are added to nominal or verbal base word in order to obtain noun and verbal forms. The process is however more than mere addition as there may occur certain morphophonemic changes in the base as well as in the affix in the process resulting in a usable form. The verb forms are derived from verb-roots or *dhātus*. These *dhātus* are encoded with the core meaning of the verb. These can be primitive[11] or derived[12]. Primitive verb-roots, which are around 2000 in number, have been listed in a lexicon named *dhātupāṭha*. They are divided in 10 groups called *gaṇas*. All the verb-roots of a group undergo somewhat similar inflectional process. Derived verb-roots may be derived from primitive verb-roots or from nominal forms. Prefixes also play an important role as they can change the meaning of a verb root. These roots then have to undergo various inflectional suffixes that represent different paradigms. In this process, the base or root also gets changed. The chart given on the next page gives an overview of Sanskrit verb roots.



8.1 Derived Verb-roots

8.1.1 those derived from verb-roots

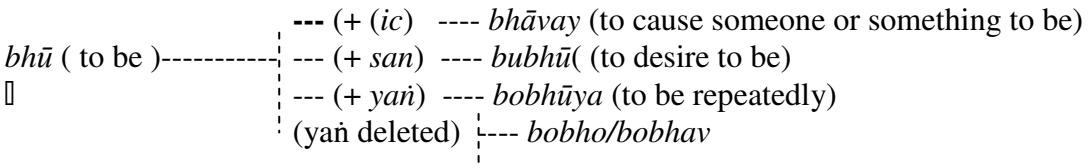
- **Causals *(ṇijanta)* -** The causals are formed by adding affix *ṇic* to a primitive verb root. They convey the sense of a person or thing causing another person or thing to perform the action or to undergo the state denoted by the root.

- **Desideratives *(sannanta)* -** Desiderative of a primitive verb root is formed by adding affix *san* to it. It conveys the sense that a person or thing wishes to perform the action or is about to undergo the state indicated by the desiderative form. Any basic verb-root or its causal base may have a desiderative form.

- **Frequentatives *(yaṅanta)* -** Frequentative verbs import repetition or intensity of the action or state expressed by the root from which it is derived. They can be of two types -
    - *Ātmanepada* Frequentative (*yaṅanta*) – affix *yaṅ* is added
    - *Parasmaipada* Frequentative (*yaṅluganta*) – affix *yaṅ* is added but deleted

[11] *bhūvādayo dhātavaḥ* (*Pāṇini* 1/3/1)
[12] *sanādyantā dhātavaḥ* (*Pāṇini* 3/1/32)

An illustration is given below of formation of derived verb-roots from a primitive verb root *bhū*.



These derived verb-roots, however, undergo similar operations, with some specifications, to form verb forms.

8.1.2 those derived from nominal words

A large number of Sanskrit verb-forms can be derived from nominal words. These are known as *nāmadhātus* (denominatives). Taking a nominal word as head, various derivational suffixes are added to these to form nominal verb-roots. The sense conveyed by the nominal verb root depends upon the suffix added to it. Yet, denominatives commonly import that a person or thing behaves or looks upon or wishes for or resembles a person or thing denoted by the noun. These denominatives, however, can be innumerable as there is no end to nominal words in Sanskrit.

8.2 Process of formation of Sanskrit verb forms

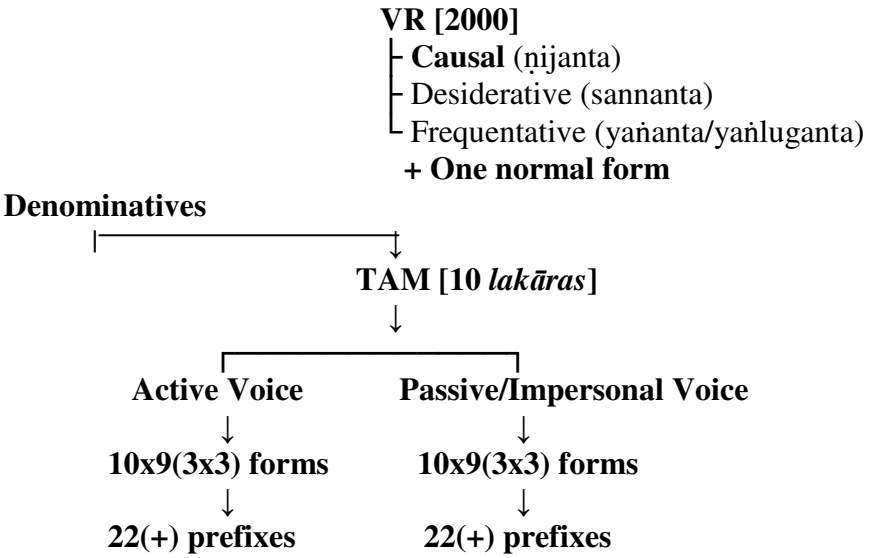
A Sanskrit verb root may take various forms in different inflectional paradigms. Sanskrit has ten *lakāras*, i.e. four moods (Indicative, Imperative, Optative, and Subjunctive) and six tenses (Present, Imperfect, Perfect, Distant Future, Future and Aorist). The *lakāras* are named in C-V-C format. The first consonant *l* signifies that the suffix has to be replaced by *tin* terminations further. The vowels *a, i, o, u, e, o, r* distinguish one *lakāra* from another. Last consonant, either *ṭ* or *ṅ*, signifies different operations. These *lakāras* are added to the root, as primary suffixes, so that it denotes a meaning in the particular tense or mood indicated by that particular *lakāra*.

Verb inflectional terminations or conjugational suffixes are 18 in number. These are divided in two groups – *Parasmaipada* and *Ātmanepada*, each having 9 affixes – a combination of 3 persons x 3 numbers. Thus each of the 18 terminations expresses the voice, person and number. A verb is conjugated in either *pada,* though some of the roots are conjugated in both. For each different *lakāra,* a root is affixed with these 9 terminations in a single *pada*. Again, there are three voices- Active, Passive and Impersonal. Transitive verbs are used in the Active and Passive voices while intransitive verbs are conjugated in the Active and Impersonal voices. The 18 inflection terminations are basically replacement of the *lakāra* or primary suffix. According to Pāṇini, when a *lakāra* is added to a root, it is replaced by 18 terminations. Thereafter, one of the18 remains to create a verb form.

For each separate *lakāra*, the 18 *tin* terminations are replaced by other forms, an illustration of the replacement technique of Pāṇini. Thus *ti, tu, tā, t* etc. are the various replacements of same affix tip in the environment of different *lakāras*.

Then we have certain characteristics *(vikaraṇa)* inserted between the root and the termination. This characteristic can vary according to *lakāra* or the class of the verb root. For four of the *lakāras*, we have *śap* as a characteristic – only for four *gaṇas*.

Addition of one or more of 22 prefixes (*upasargas*) to verb roots can result in more variety of forms. Derivative verb roots, both derived from verb roots as well as nominal words, also follow the same process to form verb forms. There can be some specific rules and exceptions in some cases. The following tree gives a rough estimate of all the possible verb-forms of Sanskrit.[13].



The verb roots of different *gaṇas* adopt certain terminations when *tiṅ* affixes are added to them. Consequently, the verb roots of these classes form verbal bases ending in *'a'*. The *tiṅ* affixation also influences the verb root and it undergoes several morpho-phonemic changes, for example, having *guṇa* operation on the end vowel. The verb root can adopt certain more operations resulting in the final verb-form.

bhū + ti(p)
↓
bhū + (ś)a(p) + ti (infixation of characteristic)
↓
bho + a + ti (penultimate vowel gunated)
↓
bhav a ti (*ayādi* sandhi)

As shown in an example, when suffix *tip* is added to verb-root *bhū*, we obtain *bhavati* as the final verb form. This *bhavati* can be analyzed in *bhav* + *a* + *ti*. Here *bhav* is the prepared verbal base whereas *a+ti* is the combination of 'characteristic + conjugational affix.' This can be cited as a common analysis of most Sanskrit verb forms. The verbal base of a verb root remains same in all its forms whereas the second combination is common for almost all the roots of a single *gaṇa.* The analysis applies to the first category of derived verb roots as well.

[13] **Mishra Sudhir K., Jha, Girish N.,** 2004**, *Identifying Verb Inflections in Sanskrit morphology*,** In proc. of SIMPLE 04, IIT Kharagpur, pp. 79-81.

9. Analysis of Sanskrit verb forms

9.1 Strategy for regular verb forms

The simplest strategy for regular verb forms can be to store all the possible forms of all the verb roots in any structured form. But given the enormity of Sanskrit verb-roots and the multiplicity of inflectional paradigms, this approach is far from being practical. A better approach may be arrived at by following the analytical method.

As illustrated above, Sanskrit verb forms are a blend of multiple morphemes which contain relevant information. Analytically, we can say that the first element is the conjugational affix that remains at the end of every verb form. These affixes have encoded information of *pada* (though it is determined by the root), *lakāra*, person and number. Thus terminations can serve as the most important thing to let us know about the paradigm information of any verb form. They can be a tool to identify a verb form in a given text. The terminations, as they are basically replacements of 18 original *tiṅ* affixes in different *lakāras*, differ among themselves according to the *lakāra*. However, in each *lakāra* they are similar for all the verb roots of various groups, leaving some exceptions. So, *ti* can be used to identify any verb form of present tense in *parasmaipada*. But some terminations can vary among themselves for a group of *gaṇas*. Then again, the terminations may be changed due to morphophonemic environment, *tā* affix of *luṭ lakāra* changing to *ṭā* with roots like *yaj*.

Further left we have the remaining morphemes of the various characteristics and increments inserted between the verb root and terminations, in the process of their formation explained above. So, *bhvādigaṇa* verb forms, in conjugational *lakāras*, have '*a*'- a result of *śap* characteristic; *svādi* roots have *no*, *nu* or *nv* - all of them remaining morphemes of *śnu*. Some roots like that of *adādi* have no such characteristic sign infixed in them.

Then we have the modified stem of the verb root at the right end of the verb form. The modification can be that of *guṇa*, *vṛddhi* or any other. Generally a root adopts a common stem in all the forms for both the *padas* in conjugational *lakāras*. So, *bhav* is the stem for all the *parasmai* forms in the conjugational *lakāras*. But there are exceptions to it to the extent that four or five types can be found among nine forms of a single *lakāra-pada*.

Here, the first morpheme- the *tiṅ* termination is common among all the verb forms of a particular *pada-lakāra*-person-number combination. Second constituent- the characteristic (existing in the form of its remaining morpheme) and increments inserted in between may differ, yet being almost the same in a particular group. The third constituent- the modified verb-root is particular in the strict sense. In the analysis, the recognition of the *tiṅ* termination will identify a word as a verb form and find out its *pada-lakāra*-person and number. The second morpheme can, in many cases, be helpful to recognize the *gaṇa* of a particular root because the characteristics in a *lakāra* are determined by the *gaṇa* that the root belongs to.

Thus the core of the analytical approach is that each *tiṅanta* verb form can be analyzed to form a unique combination of verbal stem + *tiṅ* termination; and we store both of these constituent parts in separate tables. So, when it is to be analyzed, its constituent morphemes are recognized and identified with the help of pre-stored structured data.

An example of this strategy is shown in the table given below. The first column demonstrates the representative verb root of each class. When the verbal affix *tip* is added to each of them, it undergoes certain morphological operations and results in the usable *tiṅanta* verb form listed in the second column. This is the forward Pāṇinian process. The next column demonstrates the reverse Pāṇinian approach for analysis of verb forms. In the second column every form has *ti* ending. When we remove this ending along with the conjugational affix, we obtain the storable verbal base. Every verb form can be analyzed similarly in ending and remaining verbal base.

Verb-root		**Verb-form**		**Verb-base**
bhū		*bhavati*		*bhav (-ati)*
ad		*atti*		*at(-ti)*
hu		*juhoti*		*juho(-ti)*
div		*dīvyati*		*dīvy(-ati)*
su	+ *ti* →	*sunoti*	- *ti* →	*sun(-oti)*
tud		*tudati*		*tud(-ati)*
chid[14]		*chinatti*		*chinat(-ti)*
tan		*tanoti*		*tan(-oti)*
krī		*krīṇāti*		*krī(-ṇāti)*
cur		*corayati*		*coray(-ati)*

9.2 database for verb analysis

The database tables given below demonstrate the structure of storage of all possible verbal bases of a verb root. As a sample data, five verb roots of different *gaṇas* have been taken -

table: verb bases

root	**gaṇa**	**pada**	**seṭ/ aniṭ/ veṭ**	**lakāra**	**Verbal Bases**				
					Regular	**Causal**	**Desider.**	**Frequentative**	
								Ātmane	**Parasmai**
bhū	*bhvādi*	*parasmai*	*seṭ*	*laṭ/loṭ/vli*	*bhav,*	*bhāvay*	*bubhūṣ*	*bobhūy*	*bobhav ,bobho*
				liṭ	*babhūva*	*bhāvayāñ/m*			
				laṅ	*abhav*	*abhāvay*			
				ali	*bhū*	*bhāv*			
				luṅ	*abhū*	*abībhav*			
ad	*adādi*	*parasmai*	*seṭ*	*laṭ/loṭ/vli*	*ad,at*	*āday*	*jighats*	-	
				liṭ	*jaghāsa, jaghasa jakṣa,āda*	*ādayñ*			
				laṅ	*ād,āt*	*āday*			

[14] *rudh* shows an exceptional behaviour, so *chid* has been taken.

				ali	*ad*	*ād*			
				luṅ	*aghas*	*ād*			
hu	*juhotyādi*	*parasmai*	*aniṭ*	*laṭ/loṭ/vli*	*juho, juhu,juhv*	*hāvay*	*juhūṣ*	*johūy*	*joho*
				liṭ	*juhavāñ/m, juhāva, juhuv*				
				laṅ	*ajuho,ajuhu*				
				ali	*hū*				
				luṅ	*ahau*				
div	*divādi*	*parasmai*	*seṭ*	*laṭ/loṭ/vli*	*dīvy*	*devay*	*dideviṣ*	*dedīvya*	-
				liṭ	*didev,didiv*				
				laṅ	*adīvy*				
				ali	*dīv*				
				luṅ	*adev*				

The second table illustrates the structure of the storage of verbal terminations of five *gaṇas* in both *padas* for *laṭ lakāra*. More than one termination in a single box has been separated.

table: verb affixes

lakāra	**pada/gaṇa**		**I per.**			**II per.**			**III per.**		
			Sing	**Dual**	**Plu.**	**Sing.**	**Dual**	**Plu.**	**Sing**	**Dual**	**Plu.**
laṭ	**parasmai**		*ti/ ati/ oti/ īti*	*taḥ/ ataḥ/ utaḥ*	*nti/ anti /vanti*	*si/ṣi/ asi/ aṣi/o ṣi/osi*	*thaḥ/ athaḥ/ uthaḥ*	*tha/ atha/ utha*	*mi/ āmi/ omi*	*vaḥ/ āvaḥ /uvaḥ*	*āmaḥ / maḥ/ umaḥ*
	ātmane		*te/īte /ūte ute/ īte*	*ete aate/ īyāte/ uvāte*	*ante / ate/ īyate/uvate*	*se/ īṣe/ uṣe ase*	*ethe/ āthe/ īyāthe/ uvāthe*	*adhve / īdhve / udhve*	***e/*** *īye/ uve*	*āvahe/īvahe/uvahe*	*āmahe/īmahe/umahe*
laṅ	**parasmai**		*t/at*	*tām/ atām*	*n/an*	*ḥ/aḥ*	*tam/ atam*	*ta/ ata*	*m/am*	*āva*	*āma*
	ātmane		*ata*	*etām/*	*anta*	*athāḥ*	*ethām*	*adhvam*	*e*	*āvahi*	*āmahi*

luṅ	**parasmaipadi**		*īt*	*iṣṭām*	*iṣuḥ*	*īḥ*	*iṣṭam*	*iṣṭa*	*iṣam*	*iṣva*	*iṣma*
	ātmanepadi		*iṣṭa*	*iṣātām*	*iṣata*	*iṣṭhāḥ*	*iṣāthām*	*idhvam/idhvam*	*iṣi*	*iṣvahi*	*iṣmahi*

For identification and analysis, the suffixes should be given a descending character sequence. So *ti* of *iṣyati* in *bhaviṣyati* cannot create any ambiguity.

10. Problems and possible solutions

- Verb forms which have no mark of termination left in the end are difficult to identify with the proposed module. So *bhava*, *babhūva* and other alike forms are to be stored separately.

- Some forms which are not *tiṅanta* but are similar to them like *bhavati, bhavataḥ* which are singular and dual of *bhū* in present *parasmai* third person, and also locative singular and ablative/relative singular of nominal root *bhavat.* The resolution of ambiguity here will demand involvement of semantic and syntactic analysis.

- Denominatives are formed by deriving verbal base from nominal base with the help of affixes such as *kyac, kāmyac, kyaṣ, yak, kyaṅ* etc. and then adding various verbal terminations to these verbal bases. Thus they undergo same operations and processes as regular and derived verb forms. Still there analysis is difficult due to two reasons. Firstly, nominal bases can be innumerable and thus the above stated strategy of storing the bases of all the nominative verbal bases is impossible in this case. One has to follow the rule based analytical approach. The verbal terminations can be determined with the help of affix tables as denominatives are affixed with same verbal affixes. The remaining base, however, has to be analyzed in order to infer the nominal base of that denominative. As there are some common rules to derive the verbal stem from nominal base, we can develop an analysis rule based module to identify the nominal root and can find its meaning with the help of a lexicon.

- Addition of prefixes to the verbal bases may cause morphological as well as semantic change to a verbal form. To identify one or more prefix in a verbal form, all the prefixes have to be stored in a database table along with their meaning. The system will have to check the input verbal form from left to identify single or combined prefixes. A prefix can happen to completely modify the meaning of a verb. So, creating a separate table that stores the altered meanings of various roots, when affixed with certain prefix, may be helpful in this case.

11. Conclusion

The proposed strategy to analyze Sanskrit verb forms in given text is different from existing works in many ways. It works with a reverse Pāṇinian approach to analyze *tiṅanta* verb forms into there verbal base and verbal affixes. The methodology accepted to create database tables to store various morphological components of Sanskrit verb forms is clearly in line with the well defined and structured process of Sanskrit morphology described by Pāṇini in his *Aṣṭādhyāyī*. It comprehensively includes the analysis of derived verb roots also. Even in the case of verb roots derived from nominal words, the table of affixes can provide assistance in order to separate the denominative verbal base from the verbal terminations.

References

1. BHARATI, AKSHAR, VINEET CHAITANYA & RAJEEV SANGAL, 1991, *A Computational Grammar for Indian languages processing*, Indian Linguistics Journal, pp.52, 91-103.
2. BHARATI, AKSHAR, VINEET CHAITANYA AND RAJEEV SANGAL, 1995, *Natural Language Processing: A Pāṇinian Perspective*, Prentice-Hall of India, New Delhi.
3. CARDONA, GEORGE, 1967, *Pāṇini's syntactic categories*, Journal of the Oriental Institute, Baroda (JOIB) 16: 201-15
4. CARDONA, GEORGE, 1988, *Pāṇini: his work and its tradition (vol.1)*, Motilal Banarasidas, Delhi,
5. CARDONA, GEORGE, 2004, *Some Questions on Pāṇini's Derivational system*, procs of SPLASH, iSTRANS, Tata McGraw-Hill, New Delhi, pp. 3
6. JURAFSKY DANIEL AND JAMES H. MARTIN, 2000, *Speech and Languages Processing,* Prentice-Hall, New Delhi
7. EDGREN , A. H., 1885, *On the verbal roots of the Sanskrit language and of the Sanskrit grammarians*, Journal of the Americal oriental Society 11: 1-55.
8. HUET, G'ERARD, 2003, *Towards Computational Processing of Sanskrit, Recent Advances in Natural Language Processing*, Proceedings of the *International Conference ICON*, Mysore, India
9. JHA, GIRISH N. et al., 2006, *Towards a Computational analysis system for Sanskrit*, Proc. of *first National symposium on Modeling and Shallow parsing of Indian Languages* at Indian Institute of Technology Bombay, pp 25-34
10. JHA, GIRISH N, 2003 *A Prolog Analyzer/Generator for Sanskrit Noun phrase Padas*, Language in India, volume-3,
11. JHA, GIRISH N, 2004, *Generating nominal inflectional morphology in Sanskrit*, SIMPLE 04, IIT-Kharagpur Lecture Compendium, Shyama Printing Works, Kharagpur, WB. Page no. 20-23.
12. JHA, GIRISH N., 1993, *Morphology of Sanskrit Case Affixes: A computational analysis*, M.Phil dissertation submitted to J.N.U., New Delhi
13. JHA, GIRISH N., 2004, *The system of Pāṇini*, Language in India, volume4:2,
14. JOSHI, S. D., 1962, *Verbs and nouns in Sanskrit*, Indian linguistics 32 : 60- 63.
15. KAPOOR, KAPIL, 1985, *Semantic Structures and the Verb: a propositional analysis*, Intellectual Publications, New Delhi

16. MISHRA SUDHIR K., JHA, GIRISH N., 2004, *Identifying Verb Inflections in Sanskrit morphology*, In proc. of SIMPLE 04, IIT Kharagpur, pp. 79-81.
17. MISHRA, SUDHIR K & JHA, GIRISH N, 2004, *Sanskrit Kāraka Analyzer for Machine Translation*, In SPLASH proc. of iSTRANS, Tata McGraw-Hill, New Delhi, pp. 224-225.
18. MITKOV RUSLAN, *The Oxford Handbook of Computational Linguistics*, Oxford University Press.
19. NARAYAN MISHRA, 1996, (ed). *Kāśikā of Pt.Vāmana and Jayāditya* ,Chaukhamba Sanskrit sansthan, Varanasi
20. NOOTEN, B. A. VAN, *Pāṇini's replacement technique and the active finite verb*, University of California, Berkeley.
21. SHARMA, RAMA NATH, 2003, *The Aṣtādhyāyī of Pāṇini*, Munshiram Manoharlal Publishers Pvt. Ltd., Delhi.
22. SHASTRI, BHEEMSEN, *Laghusiddhāntakaumudī (1st part)*, Bhaimee Prakashan, 537, Lajapatrai Market, New Delhi
23. SHASTRI, SWAMI DWARIKADAS, 2000, 'The Mādhavīya Dhātuvṛtti by Sāyaṇacārya', Tara Book Agency, Varanasi.
24. SUBASH & JHA, GIRISH N., 2005, *Morphological analysis of nominal inflections in Sanskrit*, presented at Platinum Jubilee International Conference, L.S.I. at Hyderabad University, Hyderabad, pp-34.
25. SUBASH, 2006, *Machine recognition and morphological analysis of Subanta-padas*, M.Phil dissertation submitted to J.N.U., New Delhi.
26. UPADHYE, P.V., 1927, *Dhāturūpacandrikā*, Gopal Narayen & Co, Bombay.
27. WHITNEY, W.D., 2002, *History of Sanskrit Grammar*, Sanjay Prakashan, Delhi.

Web references:

- IIIT, Hyderabad, http://www.iiit.net/ltrc/Publications/Techreports/tr010/anu00kbcs.txt (accessed: 22nd April 2007).
- Peter M. Scharf and Malcolm D. Hyman, http://sanskritlibrary.org/morph/ (accessed: 12 August 2006).
- Huet's site http://sanskrit.inria.fr/
- Prajna system, ASR Melcote, http://www.sanskritacademy.org/Achievements.htm
- Aiba, Verb Analyzer for classical Sanskrit, http://wwwasia.human.is.tohoku.ac.jp/demo/vasia/html/
- Desika, TDIL, Govt. of India, http://tdil.mit.gov.in/download/Desika.htm
- RCILTS, JNU, http://rcilts.jnu.ac.in
- Shabdabodha, ASR, Melcote, http://vedavid.org/ASR/#anchor2