ISSUES IN COMPUTATIONAL PROCESSING OF SANSKRIT

Veda Vāridhi, Veda Vijñana Śiromaṇi, Vaidika Bhūṣaṇam P. Rāmānujan
C-DAC, 2/1, Brunton Road, Bangalore - 560 025, INDIA.

Abstract

*We narrate our experiences at C-DAC in the past eight years relating to the development of a preliminary Natural language Understanding (NLU) System for Sanskrit, **DEŚIKA**, Computational Rendering of Pāṇinian Grammar and Creation of Ṛgvedic Saṃhitā database.*

Also discussed are the issues and challenges involved in the effort to revive the language in a relevant way to current times and the entire scope of the endeavour for success.

Introduction

Sanskrit is one of the ancient languages of India, containing a very rich treasure of literature encompassing every possible sphere of human thought, skill or communication ranging from philosophy to folk arts and crafts. Once considered Deva-bhāṣā, the tongue of the Gods, it has currently fallen out of practice and use (perhaps the reason for its inclusion under `ethnic` group ?).

Few serious and concerned researchers who see the merit in the language and its literature, its beauty and structure etc. would like it restored to its pristine glory within their lifetimes.

Standard to represent the character set

One of the first requirements is to have a way of representing knowledge contained in Sanskrit literature (Vedic and Classical) through I/O devices for data entry, visual display/printing etc. in various Indian Scripts including Roman and then to have facilities for dealing with meanings through appropriate schemes. ISCII Standard (IS 13194:1991) largely satisfies the former requirement. Annex - G of this Standard deals with Vedic accent-marking for Ṛg, Yajus and Atharva Veda-s. Inclusion of Grantha Script as an alternative to Dēvanāgarī for Sanskrit (including Vedic) and provision for Sāma Vedic accent marking scheme for Saṃhitā, Pada-pāṭha, Rahasya/Ūha Gāna, Brāhmaṇa, Upaniṣad etc. is currently under development.

Knowledge Base = Data Base + Rule Base

Further, Knowledge Bases are to be built for the entire range of literature (both Vedic and Classical) which require facilities for Data Processing besides Word Processing. These are also available in the C-DAC's GIST range of products as both hardware and software solutions on a variety of platforms, media and operating Systems. For details pl. ref. http://www.cdac.org.in/~gist/.

Range of literature - Definition of Scope

Spoken and written forms of accented and plain texts in prose, poetry and mixed kinds of terse aphorisms, prescriptive and descriptive types of works, technical treatises, commentaries, expositions, translations, polemic works, theoretical formulations or abstractions of diverse domains commencing from Vedic texts to folk arts and crafts are part of the rich literature. For a detailed road map of ancient Indian sciences, ref. the author's URL at http://www2.csa.iisc.ernet.in/~ramanuja/map.html. Efforts put in to explore the feasibility and provide assurance for handling real-world situations in many of these areas with the kind of effort put in and its effect for the field is enumerated below.

All intellectual pursuits are classified into three levels of learning, viz. **Anubhava (Upāsanā)** - Experience (consciousness), **Jñāna** - Knowledge (pure Science) and **Kauśala** - Skill (applied science). Accordingly, *Śāstra*-s dealing with these are called **Parā Vidyā, Aparā Vidyā** and **Kalā.** These also have four phases each, viz, Learning, Reflecting, Practice and Propagation as mentioned in : अधीति बोधाचरण प्रचारणैः दशाश्चतस्रः प्रणयन्नुपाधिभिः । चतुर्दशत्वं कृतवान् कुतः स्वयं न वेद्मि विद्यासु चतुर्दशस्वयम् ।।

Parā Vidyā or Brahma Vidyā, numbering 32 and Nyāsa Vidyā, deals with meditation and self-realisation on the basis of Upaniṣad portion of Veda-s (scriptural texts). These are elaborated in Brahma Sūtras of sage *Vyāsa*.

The **Aparā Vidyā**-s (**Śāstra**-s) are classified into **FOURTEEN** subjects of study or **Vidyā-sthāna**-s (or **Dharma-sthāna**-s) which include the four **Veda**-s (scriptures), six **Vedāṅga**-s (Vedic auxiliary sciences) and four **Upāṅga**-s (supplementary subjects). The aṅgas and upāṅgas of the Vedas are to be used for the proper interpretation of the Vedic texts which alone are the sole repository of all knowledge leading to various attainments (material and spiritual).

Śāstra-s - description

The Veda-s are named **ṚgVeda** (in the form of laudatory hymns of deities etc.), **Yajur Veda** (describing sacrificial/ritualistic aspect of the use of Vedic Hymns), **Sāma Veda** (musical recitation of the Hymns in Sacrifices etc.) and **Atharva Veda** (propitiatory and mundane aspects like mental and physical health, warfare, magic etc).

The six **Vedāṅga-s** (*lit.* limbs of Veda-Puruṣa) are **Śikṣā** (Science of Phonetics of syllables and pronunciation) - *Nose*, **Vyākaraṇa** (Grammar or the Science of words) - *Mouth*, **Chandas** (Science of Prosody/ Metrics) - *Feet*, **Nirukta** (Science of Etymology or literals and their Exegesis) - *Ears*, **Jyautiṣa** (Science of Astronomy) - *Eyes* and **Kalpa** (Ritual Directory specifying the practical use of Vedic texts in sacrifices) - *Hands*.

The **Upāṅga-s** are **Mīmāṃsā** (Science of Epistemology and Discourse), **Nyāya** (Metaphysics, Logic and Syllogism or Science of Sentences), **Purāṇa** (Epics which illustrate and amplify Vedic Thoughts) and **Dharma Śāstra** (Moral Code of Rectitude or Behaviour).

It is clear from the above that the entire gamut of knowledge-related issues owe their origin to Vedic texts alone in the Indian Civilisation, Culture and Philosophy which collectively is our great Heritage.

Details of the above :

Salient Features/Details - Veda-s :

Veda-s are four in number, usually likened to a tree with four major trunks and branches (Śākhās) further in each trunk. Among these, Ṛg Veda had 27 Śākhās, Yajur Veda initially divided into two major sub-types as Śukla (white) and Kṛṣṇa (black) had 15 and 86 Śākhās respectively, Sāma Veda had 1000 Śākhās and Atharva Veda 9 Śākhās.

*Currently, only two Śākhās of Ṛg Veda (there too, one of them is not very significantly different from the popular one), two in Śukla Yajur Veda, four in Kṛṣṇa Yajur Veda, **three in Sāma Veda (of which one has followers in single digits), and two in Atharva Veda (here also the number of reciters is in single digits)** are surviving in tradition. The various Vedas and their Śākhās being recited by various schools depend also upon geographical locations with the highest concentration in South India (Tamil nadu, Andhra Pradesh, Karnataka, Maharashtra and Kerala) and more sparse distribution elsewhere.*

In every Vedic Śākhā, there are four types of texts called the **Saṃhitā** (to which portion only Pada-pāṭha, Vikṛtis and Varṇa-krama apply), **Brāhmaṇa, Āraṇyaka** and **Upaniṣads**. There are special Vedic grammar rules for each Śākhā called 'Prātiśākhya' and phonetic rules known as 'Śikṣā'. There are also other 'Lakṣaṇa Grantha's which deal with accent combinations (Sandhi), Vikṛti formation etc. for a particular Śākhā.

Work content

In terms of quantity, any Vedic Śākhā would typically contain about one lakh words of five syllables per word on an average, about 1.5 lakh words of Brāhmaṇa, Āraṇyaka and Upaniṣad, similarly. With the Pada-pāṭha and various auxiliary works like Prātiśākhya, Śikṣā etc. amounting to another 1.5 lakh words, it would come to approximately 7.5 MB of text per Śākhā. For the 13 Śākhās now available as aforesaid, the total Vedic texts would amount to 98.5 MB of info. If vikṛtis are also considered, this figure would grow to 20 MB per Śākhā and Varṇa-krama 1.25 MB per Śākhā. (276.5 MB additionally).

[As it is manually impossible to print this mass of all the Vedas and their numerous Śākhās in all these forms and Varṇa-krama, the texts in Saṃhitā form (and to some extent Pada-pāṭha) are only available in print. The advent of Information technlogy can help make these information available on demand.]

There are twenty-six parameters for each Vedic syllabic definition. Given below is the typical Vedic character set with these details. As an illustration, for the vowel 'a' , consonant 'ka', and the accent 'udātta', the descriptions are as follows (respectively) :

'नाद ध्वनि जनित नात्युपसंहृत हनु स्थान तथाभूतोष्ठ करण संवृत प्रयत्न एक मात्रिक वायु देवताक ब्राह्मण जाति ह्रस्व संज्ञा सहित **अकार**', 'श्वास ध्वनि जनित हनुमूल स्थान जिह्वामूल करण स्पृष्ट प्रयत्न अर्ध मात्रिक पराङ्गभूत वायु देवताक ब्राह्मण जाति प्रथम स्पर्श संज्ञा सहित **ककार**', 'भूमि देवताक ब्राह्मण जाति सात्विक गुण सहित तर्जिन्यङ्गुलिमध्य रेखान्यास योग्य अध्येतृदेह दैर्घ्य ध्वनि काठिन्य कण्ठाकाश कृशत्व जनित अजारुततुल्य गन्धार स्वर हेतुभूत मूर्धस्थानोद्भव **उदात्त** स्वर सहित'.

The recess between different components of a character, between characters of a word etc., are also explicitly mentioned. This is the "complete" description of every unit of a Vedic syllable and happens to be a definition or specification for Vedic texts, where all aspects like literal, phonetic, grammatical, physiological, conceptual, spritual and philosophical are covered and this is the finest text-form safeguarding mechanism devised in the oral tradition. We have now created a computational database of these intricate details for the normal and special cases of Taittirīya śākhā, to begin with. This is not much different for other śākhās as well.

Vedāṅga-s :

1. **Śikṣā** of **Pāṇini, Bhāradvāja, Vyāsa, Yājñavalkya, Nārada, Āpiśali, Pāri, Kauṇḍinya** etc. deals with phonetics for Vedic intonation and accents, pronunciation and certain special features of Vedic grammar etc, in Sūtra and verse forms. **Śikṣā** is first of the six Vedāṅga-s, likened to the nose of Veda-s. Many sages have described Phonetics, and as such, different śikṣā works are attributed to them. Some of these are common for all Veda-s while some are specific to certain Veda/Śākhā. *Pāṇinīya śikṣā* is in the form of sixty verses.

According to Pāṇini, the purpose of the Science is to clarify the proper pronunciation of utterances (vācaḥ uccāraṇavidhi vyaktīkaraṇam - pā. śi. 2). As per *Taittirīyāraṇyaka,* Śikṣā deals with the pronunciation of the character set of Sanskrit language at varṇa, svara, mātrā, bala, sāma and santāna levels. What these signify is further explained.

varṇaḥ : literals (vowels, consonants etc.) beginning with a. These total 63 or 64 (pā.śi. 3) twenty-one *svaras* (Vowels), twenty-five sparśas (stops), eight beginning with *ya* (four semi-vowels and four fricatives), four yamas (doubled characters under specified circumstances - explained in prātiśākhyas), anusvāra, visarga, jihvāmūlīya, upadhmānīya, ḷakāra and ḹkāra pluta (pā.śi. 4-5).

svaraḥ : accent or intonation of three basic types, viz., udātta (acute), anudātta (grave) and svarita (circumflex).

mātrā : duration of utterance - hrasva (short), dīrgha (long) and pluta (prolated) for vowels.

balam : covers places of origin (eight) of sound and effort of articulation (internal and external).

sāma : balanced way of pronunciation (pā.śi. 31) having acceptable qualities (pā.śi. 33, 36-37) and without defects (pā.śi. 32, 34-35).

santānaḥ : Euphonic combinations (including their absence at specified situations).

The efficacy of these being in chanting, we should direct all available state-of-the-art devices towards this task, which is also technologically challenging for analysis etc. Today, there may not be proper awareness even, as to handling many of these issues like study of accent-related meaning changes etc., which hold the key to understanding spoken languages.

2. **Aṣṭādhyāyī** of sage **Pāṇini**, a Vedāṅga likened to the mouth of Veda Puruṣa, contains well-stuctured grammar rules for sanskrit language consisting of about 4000 sutras in 8 chapters and 32 quarters (in all, having only about 60,000 characters!).

Pāṇini's Aṣṭādhyāyī deals with word-level aspects like rules for formation of valid word-forms, description of their structure and rules for their usage. This covers simple, compound and concatenated word-formation of various types, parts of speech like nouns, verbs, attributives, relational functors, governance clauses etc. in scriptural and literary language, parameters like origin of sound, internal and external effort, phonetic classification, accents etc. at substrate and affix levels for all words. The substrate could be nominal stem or base for nouns and verbal roots for verbs. Affixes include prefixes (like upasarga for verbs), in-fixes (certain taddhitas for nouns, conjugational ones in certain classes of verbs etc.) and suffixes (like Sup for nouns, 'Tiṅ' etc. for verbs and so on). The process of derivation of valid word-forms becomes thus clearly defined.

3. **Chandas Sūtras** of sage **Piṅgaḷa**, likened to the feet of Veda Puruṣa, deals with Prosody, metrics and other aspects of prose and poetry applicable to Vedic as well as classical literature containing 286 sutras in 18 chapters. In Veda-s, metre is of various forms at syllable, triad, quarter, hemistich, hymn, collections, prose levels and of homogenous and heterogenous texts, with single, three/four or seven tonal variations in different Veda-s and their parts.

4. **Nirukta** of sage **Yāska**, a Vedāṅga likened to the nose of Veda Puruṣa, deals with epistemological, exegetical, phonological, morphological and semantic aspects of Vedic literature with a dictionary of 1773 vedic words, fully derived and explained with examples, giving out criteria for such an analysis, containing 12 chapters, 49 quarters and 3 major Sections, in prose order.

Yāska's Nighaṇṭu and Nirukta is the earliest available technical treatise ever in the history of human civilisation after Vedic texts and is hailed as the best treatment of linguistic and exegetic aspects pertaining to accented, phonetic forms of Vedic literature. Its study, owing to the rigorous efforts involved, is almost become extinct, except for a very few traditional scholars who are past their prime. Even publications of research in this important field have been very rare, only a couple of them in this century.

The Computational Study of Nirukta will benefit students of Sanskrit Etymology in particular and Linguists in general, regarding various exegetic aspects of systematic, phonetic languages like Sanskrit from the Vedic texts to current colloquial speech.

5. **Jyautiṣa** of sage **Lagadha**, a Vedāṅga likened to the eyes of Veda Puruṣa, describes astronomical aspects for Vedic rituals and sacrifices, in verse form (2 - Ṛk, Yajus - recensions with 35 and 44 verses). It is said that Veda-s are meant for Vedic sacrifices (as means of attaining the goals of life), which are to be performed at precise moments and durations specified. Hence, this science prescribes the time for various Vedic rituals and one aware of this science alone knows Veda-s.

6. **Kalpa Sūtras**, typically of sage **Āpastamba**, (Āśvalāyana, Bodhāyana, Satyāṣāḍha, Bhāradvāja, Hiraṇyakeśin etc.) is divided into 6 portions -

i) Paribhāṣā Sūtra-s being the metalanguage and conventions used in Kalpa Sūtra-s in 160 sutras in 4 quarters of 1 chapter,

ii) Dharma Sūtra-s dealing with moral rectitude, conduct and character, physical, spiritual and ritualistic aspects of human behaviour at individual and society levels, contained in 1362 Sūtra-s of 2 chapters, 11 sections and 29 sub-sections,

iii) Gṛhya Sūtra-s dealing with domestic Vedic ritualry, consisting of 405 Sūtra-s, 8 sections and 23 sub-sections,

iv) Pitṛmedha Sūtra-s describing obsequial ritualry comprising 306 Sūtra-s in 2 chapters,

v) Śulba Sūtra-s describing constructional and architectural aspects of sacrificial altars etc. involving **Vedic Mathematical principles**, contained in about 300 Sūtra-s, 6 sections and 21 sub-sections, and

vi) Śrauta Sūtra-s describing the various sacrifices mentioned in Vedas, contained in over 10000 Sūtra-s in 23 chapters and scores of sections and sub-sections.

Upāṅga-s :

1. (Pūrva) **Mīmāṃsā Sūtra-s** of sage **Jaimini**, an Upāṅga, contains 2617 Sūtra-s in 12 chapters, 60 quarters, 907 topics each giving out a maxim/rule for interpretation of Vedic texts.

1a. **Brahma Sūtra-s** of sage **Vyāsa**, an Upāṅga, have 545 Sūtra-s in 4 chapters, 16 quarters and 156 topics, devise methodology for analysing Upaniṣadic texts. This is also called as Uttara Mīmāṃsa and is counted as one branch of study alongwith (Pūrva) Mīmāṃsā.

The Mīmāṃsā Śāstra is primarily meant to devise methods for interpretation of Vedic texts, including Upanishads, and does so mainly at discourse level. Here, criteria for determination of discourse import, coherence, conflict resolution, priorities and relative strengths of various instruments of knowledge etc. are described. Theories of the process of cohering of word-meanings at sentence level are a hall-mark of this shastra, which has applications in all other shastras as well. Contextual aspects are dealt with in good detail and methodologies evolved by generalisation, to cater to different kinds of conflicting statements in a discourse.

2. **Nyāya Sūtra-s** of sage **Gautama**, an Upāṅga, dealing with sensory perception, Inference, Analogy, Verbal Testimony, various aspects of polemics, debate, syllogism, fallacies etc. in all describing 16 entities 'worth knowing about for realising the ultimate', in 528 Sūtra-s covered in 5 chapters, 10 Āhnika-s, and under 84 Prakaraṇa-s (major topics);

The Nyāya Śāstra deals with Ontological classification of things and proceeds to enumerate, define and verify their essential and typical characteristics. Instruments and objects of knowledge are dealt with in detail. The process of 'human understanding' is described and theories of valid knowledge, error, word-meaning relations, cognition, validity/fallibility etc. are postulated. The linguistic, psychological, ontological, logical and philosophical issues are elaborated besides covering the inference in great detail. Aspects of sentence-hood with necessary criteria like proximity, expectancy and compatibility are evolved and sentential import extraction explained. Thus, this shastra could be thought of as dealing at sentence-level processes.

2a. **Vaiśeṣika Sūtra-s** of sage **Kaṇāda**, also dealing with logical aspects, particularly mataphysical and material properties with reasoning, contained in 369 Sūtra-s in 10 chapters and 20 Āhnika-s;

3. **Purāṇa-s** (Epics) are 18 in number, authored by the Sage Vyāsa amd illustrate mythologically, in narrative style, the message and teachings of Veda-s.

4. **Dharna Śāstra-s** lay down the code of conduct for universal harmony and welfare for every section of society and are the operative part of Vedic wisdom.

Others : Darśana Śāstra-s

Sāṅkhya Sūtra-s of sage **Kapila**, describe about Nature, its evolutes, theory of cause and effect, metaphysics etc. in 555 Sūtras spread over 6 chapters.

Yoga Sūtra-s of sage **Patañjali**, deal with mental states, control of body and mind harmoniously and rigorous means of self- control and realisation, consisting of 195 sutras in 4 quarters of one chapter [There is another recension having 4 chapters with 4 quarters each, having about 410 sutras.].

(Nāṭya Śāstra of Bharata), Alaṅkāra Sūtra-s of **Vāmana** deal with classical literature contained in 12 chapters and about 300 sutras. Figures of Speech, Allegory etc. in prose, poetry and mixed forms, drama, dance, various adio-visual forms etc. are covered to put across sagely advice in an unobtrusive manner.

The **Kalā-s** (applied Sciences), numbering 64, cover the following : Itihāsa (history/legion), Āgama (idol worship/rituals), Nyāya (jurisprudence), Kāvya (classical literature), Alaṅkāra (Figurative speech), Nāṭaka (drama), Gāna (music), Kavitva (poetry), Kāmaśāstra (erotica), Dyūta Naipuṇya (skill with dice), Deśa Bhāṣā Jñāna (regional lingusitics), LipiŚarma (lithography), Vācana (oratory), Samastāvadhāna (concentration), Svaraparīkṣā (voice recognition), Śāstraparīkṣā (armoury/warfare), Śakunaparīkṣā (knowledge of omens), Sāmudrikaparīkṣā (physiology), Ratnaparīkṣā (gemology), Svarṇaparīkṣā (goldsmithy), Gajalakṣaṇa (elephant rearing), Aśvalakṣaṇa (horse rearing), Mallavidyā (wrestling), Pākakarma (cookery), Dohaḷa (pottery?), Gandhavāda (odour sense), Dhātuvāda (metallurgy), Khanivāda (minerology), Rasavāda (chemistry), Agnistambha (fire control), Jalastambha (staying afloat), Vāyustambha (wind control), Khaḍgastambha (tightrope trick), Vaśyā (hypnotism), Ākarṣaṇa (seduction), Mohana (mesmerism), Vidveṣaṇa (witchcraft), Uccāṭana (exorcising), Māraṇa (killing), Kālavañcana (time evasion), Vāṇijya (commerce), Paśupālana (animal husbandry, Kṛṣi (agriculture), Samaśarma (balancing),

Lāvukayuddha (fencing fight), Mṛgayā (hunting), Putikauśala (dollcraft), Dṛśyasaraṇi (occultism), Dyūtakaraṇi (dice control), Citraloha (alchemy), Pārṣāmṛt (?), Dāru Veṇu Carma Ambarakriyā (wood, bamboo, leather garment work), Caurya (theft), Ouṣadhasiddhi (medicinal powers), Mantrasiddhi (incantation powers), Svaravañcanā (mimicry), Dṛṣṭivañcanā (beguiling), Añjana (anointing), Jalaplavana (swimming), Vāk Siddhi (prophecy), Ghaṭikā Siddhi (prediction), Pādukā Siddhi (cobblery), Indrajāla (jugglery), Mahendra Jāla (magic).

Knowledge Representation Issues

Modern Knowledge-Based-Computer Systems employ Predicate Logic (*'if-then-else'* form of rules), Semantic Networks and Conceptual Dependency schemes to represent 'World Knowledge' in Computers. The correspondence of these three methods to the "**Śābda-bodha**" concept in the three branches of Śāstraic (Sanskrit) literature, viz. Nyāya, Vyākaraṇa and Mīmāṃsā has been shown [1]. The self-inference generating characteristic of Sanskrit grammar (of Pāṇini) has also been brought forth [2]. Thus, for a variety of applications like NLP (Natural Language Processing), MT (Machine Translation), CAL (Computer-Aided Learning) and Expert Systems, the Śāstraic concepts could be adopted. [References : 1. Analysis of sentences in Sanskrit and Knowledge Representation Techniques by **HR Vishwasa, RV Hudli and T Vishwanathan,** Papers presented at KRIS-86, Bangalore. 2. Knowledge Representation in Sanskrit and AI by **Rick Briggs,** AI Magazine, Spring, 1985.]

This poses the problem of how knowledge is to be represented in Computer systems. To solve this, an appropriate model of knowledge is required. For this, epistemological building blocks are needed. Śāstraic literature in Sanskrit contains exhaustive treatment of the various aspects of knowledge, *per se*, and attempts to systematise its study. Knowledge elements for word, sentence and discourse levels are enumerated, defined, examined and established as building blocks.

The tenets of the Śāstras throw sufficient light for realising a knowledge representation based on these. We could benefit from the various methodologies and principles evolved therein for analysis of verbal communication through Sanskrit language, abstracted suitably. Three distinct constituents of a knowledge representation scheme could be a Lexicon, a knowledge base and programs for analysis of the knowledge base.

The **Knowledge Base** could be the Sūtras of all the branches of learning in Śāstraic literature. Of these, the ones relevant for knowledge based systems applications are the Aparā Vidyā-s directly and hence, their tenets need to be represented in an exhaustive knowledge base for analytical purposes. A compendium all Sūtra-s called **'Sakala Śāstra Sūtra Kośa'** would have to be prepared. These tenets are either in aphorisms (a sort of terse, pithy sentence) form or as verses. Their analysis is to be attempted through an expert system for such a research. Here, the lexicon based on Amarakōśa assists semantic extraction.

These sutras are to be coded in a uniform manner, tabulated suitably with fields giving necessary details in a form compatible with computer operation. This portion is quite laborious as the subject is vast. Here, another area of fundamental research, also advantageous to undertake through sanskrit, i.e, Speech-to-text conversion, would amply complement the present research, helping the codification of the extant

knowledge base through recitation/ recording. In the same vein, Image processing research also could help by automatic character recognition and generation from written inputs, including manuscripts.

The following list details the databases built by/available with us:

A. Veda-s(4) -

1. **Ṛgveda -**

a) Saṃhitā - text (8 aṣṭaka-s of 8 adhyāya-s each/10 maṇḍala-s, 85 anuvāka-s, 1028 sūkta-s)
b) Pada-pāṭha - split-words;
c) Khilasūktas - auxiliary texts;
d) Kātyāyana's Sarvānukramaṇī - Index to hymns with seer, metre, deity and application;
e) Ṛk prātiśākhya - Vedic grammar
f) Sāyaṇa Bhāṣya - for 30 select hymns (+ links to Aṣṭādhyāyī, Yāska's Nirukta/Nighaṇṭu)
g) English Translation for 30 select hymns (by AA McDonnell and HH Wilson for 1st aṣṭaka)
h) On-line version of a sample ṚgVedic Reader.

2. **Yajurveda -**

a) Taittirīya Saṃhitā - Pañcāśadanukramaṇikā (paragraph index) 2198
b) Taittirīya Brāhmaṇa - Daśatyanukramaṇikā (Paragraph index) 2354
c) Taittirīya Saṃhitā Padapāṭha - split-words - prathama kāṇḍe prathamaḥ praśnaḥ
d) Taittirīya Prātiśākhya - Vedic grammar for KYV.
e) Jaṭādarpaṇa & Ghanadarpaṇa - works on Vikṛti-s jaṭā and ghana, with commentaries
f) Vājasanēyi Prātiśākhya - Vedic Grammar for SYV.

3. **Sāmaveda** - NIL* **(see technologies to be developed)**

4.Atharvaveda

a) Upaniṣat - excerpts

B. Vedāṅga-s (6) -

1. **Śikṣā** - Science of Phonetics

a) Pāṇini - 5 variants; prose and poetic forms; b) Bhāradvāja - with commentary; c) Kauṇḍinya; d) Āraṇyaka; e) Āpiśali; f) Candragōmi

2. **Vyākaraṇa** - Science of Grammar - **(under DEŚIKA Project of DoE)**

a) Pāṇini's sūtrapāṭha (Aṣṭādhyāyī), rules with inheritance etc.; b) Padakōśa, vocabulary; c) Gaṇa-pāṭha; d) Dhātu-pāṭha; e) Liṅgānuśāsana; f) Program to analyse rulebase - uses DEŚIKA; g) Program for sentence generation and analysis in Sanskrit - DEŚIKA

3. **Chandas** - Science of Metrics/Prosody

a) Nāṭyaśāstra - definitions; b) Vṛttaratnākara - definitions

4. **Nirukta** - Science of Etymology

a) Yāska's Nirukta - text; b) Yāska's Nighaṇṭu - text

5. **Jyautiṣa** - texts (two recensions)

6. **Kalpa** - Ritual Directory

a) Āśvalāyana's Śrauta Sūtra - excerpts from RV. Sāyaṇabhāṣya
b) Satyāṣāḍha's Śrauta Sūtra - 12th Praśna

C. **Upāṅga-s** (4)

1. **Mīmāṃsā** - Science of Epistemology

a) Pūrvamīmāṃsā - śābdabōdha udāharaṇa vākyāni - Nyāyaprakāśa
b) Uttaramīmāṃsā - Vedānta - a) Bhagavad-gītā - text

2. **Nyāya** - Science of Metaphysics

a) Gautama's Nyāya Sūtras - text; b) Tarkasaṅgraha - text; c) Tarkasaṅgrahadīpikā - part I

3. **Purāṇa-s** - Epics (samples with syntactic tagging)

a) Bhāgavata; b) Rāmāyaṇa - text; c) Mahābhārata - excerpts; d) Viṣṇu Purāṇa; e) Brahmāṇḍa Purāṇa; f) Nṛsiṃha Purāṇa; g) Harivaṃśa and h) Viṣṇusahasranāma Stotraṃ - Bhagavadguṇadarpaṇaṃ Bhāṣya of Parāśarabhaṭṭar

4. **Dharmaśāstra** - Moral code of rectitude

a) Pañcatantra (30 episodes) and b) Hitopadeśa (some excerpts) both with syntactic tagging

D. **Others**

1. Kauṭilya's Arthaśāstra - text
2. Vātsyāyana's Kāmasūtras - text
3. Varāhamihira's Bṛhatsaṃhitā - text
4. Śaunaka's Bṛhaddevatā - excerpts

E. **Classical Literature**

a) Vāmana's Kāvyālaṅkāra Sūtra - text (and with word-split)
b) Arthālaṅkāras - definitions, examples, chart etc.

We now take up each of these domains to study their special requirements.

Vedic Literature : The phonetic details (intonation markers) are the distinguishing feature of these texts which are preserved, even while writing, by the use of special (diacritic) accent-markers. While different Vedic Śākhā-s use different marking schemes, scripts also affect the marker scheme, e.g, Devanāgarī, Grantha, Kannada, Telugu, Malayālam, Bengāli etc. use different markers for denoting Vedic accents. (the author recently came across a palm-leaf manuscript of Ṛgveda Saṃhitā *even in Tulu Script with a distinct accent-marking scheme!) Type of* Vedic text, like Saṃhitā, Pada-pāṭha, Brāhmaṇa, Āraṇyaka, Upaniṣad etc. also use different schemes in certain Śākhā-s.

Audio interface development is useful and feasible based on a rule-based approach for these texts. In fact, it is our firm belief that knowledge and use of Vedic phonetic details would be vital for devising a successful Speech recognition software for Sanskrit (and many other derived Indian languages).

Broadly, Sāmaveda differs from the other three Veda-s in the range and representation scheme of accent-markers used. It has multiple symbols for single syllables quite often and has predominantly a row above the text earmarked for accent markers. (We have come across Grantha manuscripts of Sāmaveda using some accent markers below the text line also). Lateral accent-markers are also used in Gāna texts etc. which are distinct from the other Veda-s.

Roman script, popularly in use, also has Vedic accent markers for acute and circumflex of the *nitya* or *jātya* variety alone and these two makers are adequate to determine the accents of the rest of the syllables using a rulebase. However, recently, Prof. George Cardona of University of Pennsylvania, U.S.A, has employed a more elaborate scheme of accent marking in Roman Script to make things explicit (like in the case of Devanāgarī). Accent marking in Roman Script is yet to be provided for in ISCII Standard.

The desideratum

Thus, there is a need for having an automatic scheme of transliteration among Indian languages as also from Indian languages to Roman Script. This should cover the entire range of literature, Classical as well as Vedic. While ISCII Standard aims to achieve this, presently, it caters only to accent-marker schemes in Devanāgarī for different Śākhā-s of Vedic texts (in all, 29 additional characters/symbols are provided for, as

in Annex - G of the Standard) except Sāmaveda, which is currently being developed. Inclusion of Grantha Script as an alternative to Devanāgarī, for Sanskrit literature, both Classical and Vedic, is also being addressed.

The other cases mentioned earlier are to be incorporated to complete the task. As part of a DoE-funded Project, C-DAC has set out to accomplish the task comprehensively and exhaustively. It is also proposed to bring out a detailed write-up on the same, after consulting various handwritten (palm-leaf, paper manuscripts etc.) and printed editions of all the four Vedic texts in all (available) Indian Scripts, by visiting various Oriental Libraries and traditional archive sites in the country. Internet sources regarding those available abroad can also be consulted. An Authoring System for Sanskritists will also be an outcome of this Project, including a Sanskrit Wordprocessor.

The process of Intellectual Heritage preservation can be launched wherein the Vedic texts of all Vedic Śākhā-s are to be keyed in using GIST enviornment, where, accents can also be properly included. (We can even use any Indian script besides Devanāgarī). Then, using Aṣṭādhyāyī, relevant Śikṣā-s and Prātiśākhya-s etc. we can generate the other combinatorial (Aṣṭa Vikṛti) forms like Krama, Jaṭā, Ghana, Dhvaja, Rekhā, Mālā, Śikhā, Ratha etc., and also Varṇa Krama, using DEŚIKA software. Thus, a MASTER REFERENCE can be prepared for ALL existing Vedic Śākhā-s for posterity.

Research into and understanding the contents would also be greatly assisted by the DEŚIKA software, including Vedic accent analysis, with the help of the six auxiliary sciences (Vedāṅga-s).

Platform for study of these texts

DEŚIKA has been conceived as the general-purpose Sanskrit processing system capable of lexical upgradation and utility for Classical and Vedic Sanskrit applications. It is based on Grammar, Logic and Epistemological principles and has modules for Generation, Analysis, Sandhi and Vedic processing. A detailed User's Manual is browsable at the Author's website cited earlier. Facilities for Search, Pattern-based retrieval, Index, word-level Concordance etc. for Ṛgveda Saṃhitā text and Pāṇini's Pañcagrantha-s have been developed.

Development of tools, utilities and libraries is under way at C-DAC's Indian Heritage Group under DEŚIKA software. This includes facility for analysing words, simple sentences with Śābda-bōdha (Vaiyākaraṇa, Naiyāyika, Mīmāṃsaka) and explanations in english, text (corpus), splitting sandhi (elementary and with user-intervention), automatic lexical tagging, human-assisted syntactic tagging, word index, search, retrieval based on different string patterns, concordance etc. All these are based on ISCII Standard and hence, have transliteration among Indian languages and Roman as a built-in feature.

International Status on similar efforts

Various web-sites exist which have archives of electronic form of these classics to varying standards of scripts, keyboard lay-outs, Roman transliteration schemes and fonts etc. Many of these texts are avilable for browsing/downloading, notably in Indology site, Japanese, American, German and British initiatives etc.

TITUS Project at Frankfurt University, Germany has collected quite an impressive number of texts of Vedic literature in electronic form.

Our effort is for making these available to THE standard and in all Indian and many foreign scripts of users' choice.

Challenges

- In-depth study of nature of language, logic, communication/expression etc. needed
- Creation of authentic and relevant source for 'original' technology
- Development efforts to begin with indigenous ideas in NLP, Knowledge representation, Speech recognition, OCR, machine translation, NL interfaces, educational s/w etc.

New tools/technologies to be developed

The Indian Standard BIS 13194:1991 needs enhancements and/or amendments. GIST firmware also needs upgradation. Roman Transliteration needs to incorporate Vedic diacritic (intonation) markings, as also transliteration to other Indian scripts appropriately. Sāma Vedic accent markers pose an altogether different kind of a problem as they are used like "overscript" (as against superscript kind of display/printing) and data processing capability is to be ensured. To cater to the needs of South Indian traditional Vedic scholars, the Grantha script also has to be included in GIST products, which has to be developed afresh. The current multi-set Vedic Sanskrit fonts are to be made into a single-set.

- GIST firmware - revise VEDIC overlay, add Tamil-Grantha script, VEDIC accents in Diacritic Roman & all Indian scripts
- Develop scheme for Sāma Vēda data entry and fonts
- Improve Vedic fonts of ISM and include in ALP, LEAP, SDK

Utilities (Lexware) Development for :

- Text Processing - to create Index, Concordance, thesauri, Dictionaries etc.
- Search/retrieval from large texts
- Spell-check for Sanskrit
- Web authoring tool - Features comparable to SCHEMA, OCP, LEXA etc.

Benefits/returns Expected

- An authentic and highly informative content website on Indian Heritage
- Standards for creation of content for any specific system, text, mode etc.
- Large suite of tools and utilities for better access, authoring and study
- extension to all Indian languages
- many multi-lingual, multi-media CD titles planned on Heritage, culture etc.
- repository of original treatises, manuscripts and information content for posterity