Sanskrit in the age of Information Technology

Amba Kulkarni

Abstract

In this paper on the one hand we discuss how the theories of sabdabodha can contribute to the growth of Sanskrit by strengthening the bond between Sanskrit and other languages and on the other hand various dimensions in which Sanskrit should be technologically equipped so as to be able to easily accessible in the age of Information technology.

1 Introduction

It has been well observed by several scholars on various platforms in the past that while all languages are on the receiving end of the Information Technology, Sanskrit has a give-and-take relationship with the Information Technology. While the information technology is compelling the languages to be techno friendly for its survival, Sanskrit also has a potential to make fundamental contribution to the field of information technology, provided we take steps in right direction at right time.

In the following section I highlight the role of Indian Grammatical tradition and thereby the role of Sanskrit scholars in the development of computer technology for Sanskrit followed by a section on how the Indians can benefit from the Indian grammatical tradition to build better technology, to start with, for Indian languages, and later on also for foreign languages. Finally I conclude with what needs to be done highlighting various opportunities in the field of Computational Linguistics.

2 Equipping Sanskrit in the age of Information Technology

A language is techno friendly, if it allows the user to communicate with others with the help of modern electronic gadgets such as computers, mobile phones, ipads, ipods, tablets etc. Thus these devices should have

capability to identify the script with the help of Optical Character Recognisers (OCRs),

Lecture delivered at ‘NATIONAL SEMINAR on SANSKRIT-DEVELOPMENT PLAN FOR NEXT 10 YEARS’ on 7th February 2016 at Udupi.

spell-checking and grammar checking facility in the editors,

appropriate software for speech and voice recognition which can take dictation, questions and oral commands,

speech synthesizers to answer the queries,

search engines that search and produce relevant documents,

text summarisation softwares that can produce the summary of a document automatically, and

automatic translators.

Unlike other technologies such as recording and transmission of audio signals in different languages, making a language techno-friendly poses certain challenges. The technology related to recording and transmission of audio signals is independent of languages. That is, once the technology was developed, we could use it for any language in the world. The techno-friendliness of a language, on the other hand, is language dependent. And thus, while a large part of such software can be developed independent of languages, we need to fine tune some of the components or even develop special software so that we get the best results for a chosen language. If we blindly use the technology without proper adaptation or fine tuning, we may be loosing the intricate features of our language. To give an analogy, all of us who have seen the typewriters for Indian languages would certainly admit that the type-writing technology was good for languages with linear scripts such as Roman, but was not at all proper for retaining the aesthetic part of our scripts. Our scripts are basically compositional in nature and use left, right, top and bottom space of a character for joining with other characters. It is very much necessary that the technology developers be aware of various features of the language so that they do justice to it. Further, Sanskrit has an advantage of having a rich grammatical tradition which has discussed various aspects of language communication from synthesis to analysis. we should take advantage of this scientific knowledge to provide a basis for the development of language technology for Sanskrit. And hence, it is necessary that for the development of suitable language technology solutions for Sanskrit, both the technologists and Sanskrit scholars work together.

In the following I enumerate various tasks with regards to enabling Sanskrit with various technological facilities, mentioning wherever special attention is needed taking into consideration the nature of the language, script, and the presentation of literature owing to the dominant oral tradition. Some of these tasks need Government initiatives, or policy decisions, some tasks need to be undertaken as a development activity involving several young scholars and there are certain tasks which need intensive research involving senior well-established scholars.

2.1 Overcoming the script barrier

Sanskrit, though is written primarily in Devanagari, is also written in several other native scripts such as Telugu, Kannada, Gujarati, etc. There are thousands of manuscripts in non-Devanagari scripts across India. In various parts of India, we still find Sanskrit is being taught in the native scripts. This also has resulted into several websites having Sanskrit texts in different scripts. With the availability of Unicode now it is possible to transcribe these into Devanagari and vice versa. What is being said of Sanskrit is true for other Indian languages as well. We have a good number of population which is bilingual or multi-lingual, but this population might be knowing only one script. It is necessary that all the modern gadgets such as mobiles, tablets, ipads, ipods etc. be equipped with a facility to choose a language and a script. These gadgets should be equipped with an inbuilt transliteration facility to convert a text from one Indian script to another independent of the language it is written in.

An important aspect where a policy decision needs to be taken concerns with the SMS (Short Message Service). The size of the SMS in Roman script is limited to approximately 150 characters per message. But when it comes to messages in Indian languages written in Indian scripts, the maximum size of the SMS reduces to around 60-70 characters, less than half of that of Roman script. This happens because the messages are encoded with UTF-8 encoding before they are transmitted, and UTF-8 uses 3 bytes to represent one Unicode Indian language character, while it uses only one byte to represent a Roman character and punctuation marks. It is possible to overcome this limitation, if we encode our texts uniformaly to an intermediate Roman encryption before we transmit it and after receiving, decode it before display. For this to happen, every handset should have a facility of converting the Roman texts into a script of receiver's choice. This will also encourage the users to use native scripts instead of random romanisation leading to chaos. This decision would then also impose a discipline among the users, which is necessary if in future we would like to enable on-the-fly translation facilities among various Indian languages, including Sanskrit.

Several times we find it easy to write than type. Several devices are also now equipped with writing pads, in addition to the keyboards. The availability of writing pads demands a software that can convert the handwritten image into a text. Such devices are called Optical Character Recognisers (OCRs). The technology used to build OCRs for scripts with linear representation such as Roman can not be directly used for Indian language scripts, where the syllables (aksaras) are compositional in char-acter, and the characters involved undergo some changes when joined to-gether. While these softwares use Big data, and various machine learning paradigms, detection of various language related parameters, and proper combination thereof improve the performance of such softwares. There have been several efforts towards the development of OCRs for Indian language scripts. However, still we do not have an OCR that is near perfection. Institutes such as IIIT-Hyderabad, IISC, Bangalore, ISI, Kolkata, and C-DAC have been working in this area for the past few years. There are also worth mentioning efforts outside India. Oliver Hellwig's OCR is one such example, which at present has better performance for Sanskrit texts written in Devanagari.

In addition, if OCRs for various old scripts such as *Sarada*, *Grantha*, etc. are developed, they will enable semi-automatic conversion of manuscripts into texts speeding up the process, which otherwise will take several man years given the fact that the number of scholars knowing these scripts is very small as compared to the number of manuscripts in these scripts. Development of these softwares will also enable us to transliterate these manuscripts into modern Indian scripts such as *Nagar*. The reach and accessibility of these manuscripts will thus widen.

2.2 Speech recognisers and Generators

There has been very little work in the area of Speech recognisers and Speech generators in the field of Sanskrit. Sanskrit has the necessary theoretical description of phonology, needed to develop these softwares, in the texts of *Pratisakhya* and *Sikṣa*granthas. There have been some efforts in this area at IIT-M, IIIT-Hyderabad and JNU. But the work done for Sanskrit is very primitive.

2.3 E-library

Sanskrit, in addition to being the primary culture-bearing language of India, had the status of being Lingua-franca of academic debates for several centuries. It has resulted into a wide coverage Sanskrit literature ranging from scientific works in the fields of Philosophy, Mathematics, Ayurveda, Grammar to literary works such as Poems, Dramas, including mammoth epics. Sanskrit has a strong oral tradition. We have seen in the past that with the invention of various new communication media, there were efforts to make this vast orally transmitted literature over centuries accessible through these new media as well. Thus we see Sanskrit literature was made available through palm-leaf manuscripts by scribes. It is estimated that Sanskrit has around 50 million manuscripts. With the availability of print media, several manuscripts were printed in the form of books and now as digital texts in the digital era.

What has all been done so far

Several Sanskrit texts have been made available in readable text form. In India, through the Government initiative of SanskNet project involving several institutes, thousands of classical texts have been digitised. Several volunteer organisations, NGOs, private organisations, and individuals also have made Sanskrit digitised texts available online. Some of the notable efforts outside India are the efforts by GRETIL, SARIT, and Sanskrit-library. Most of these texts are also in Devanagari, and are searchable e-texts.

What needs to be done

1. Need an authentic site with authentic version
   Typically we come across more than one version of the texts on the web. Further many-a-times these texts are not reliable, since they have typing mistakes, and the details regarding the edition, version etc. would be missing. It is necessary that there be one authentic / reliable web-site where one can find ANY Sanskrit text, with all bibliographic details as well. For other classical languages such web-sites exist. For example for Old English a project such as Gutenberg has a collection of several English texts at one place. Project Perseus has made the texts in classical languages such as Greek, Latin, and Arabic available online at one place.
   The efforts of 'Bharatavani' project are in this direction.
2. Text Encoding of E-texts
   It is not enough to make the Sanskrit texts available online. For the effective use of these texts, following needs to be done.
   - Each text should carry the bibliographic information as a header / footer, providing all the details such as author, edition, publication, year of publication, etc.
   - The texts should be annotated with various information such as title, sub-titles, shlokas, references, quotations, etc. Necessary guidelines for such annotation need to be developed. The hierarchical structure of various related texts should be made clear in the representation.
   - The e-texts should be proof-read before posting on the website.
   - The e-texts may also have a sandhi-split version to enable easy searching, and understanding.
   - The compounds may also be presented with hyphenated components.
   - Such texts can further be linked to the online computational tools such as morphological analysers which show possible analyses of each word, sandhi splitters which show possible sandhi splittings, etc.
3. Displaying the Structure
   It is not just enough to make all these texts available in electronic forms, but there is a need to evolve standards to represent these texts in electronic media to avoid mis-interpretations, mis-use of these texts. The Indian textual tradition differs from the Western tradition. In Indian tradition, the texts are typically not independent, but are

typically embedded into other texts. We have various traditions of textual representations such as

(a) Commentary tradition: Under this again we have Bhaṣya, vṛtti, vivaraṇa, vyakhyana, vyakhya̦varttika, etc. Each of these should be represented faithfully.
(b) Khaṇḍana-Maṇḍana tradition
(c) Prakaraṇa granthas
(d) Nested stories as in Pañcatantra, Kadambar, etc.

Special efforts are needed

(a) to make the discourse structures more explicit
(b) to provide the complete context
(c) to evolve appropriate tags (through Text Encoding Initiative (TEI))

2.4 Manuscripts

National Mission for Manuscripts (NAMAMI) has taken initiative in digitising various manuscripts. Under this scheme, the images of manuscripts are being stored and made available to the users. Linking these images to the actual texts would be very much valuable. This can accelerate easy search of various texts in the manuscripts, at the same time, preserving the original presentation of the texts.

2.5 Dictionaries

The rich derivational morphology of Sanskrit leading to complex derived lexicon has resulted into varied representation of lexicon in the dictionaries. For example, Monier Williams' dictionary has a four tier representation of lexicon. Such a representation also makes it difficult to search a word in the dictionary. With computers, now it is easy to present the lexicon through multiple perspectives. Several dictionaries such as Monier Williams', Apte's practical dictionary, Vacaspatyam, etc. are available in e-format. Different kosas such as Amarakosa complement the dictionaries with the kind of information they provide. And hence it is also necessary to link all these kosas with modern dictionaries presenting a holistic view of the lexical item.

In an independent effort, a Sanskrit WordNet following the modern principles of WordNet has been built by IIT Mumbai. It is fully equipped with ontological information. Amarakosa has been enriched with ontological information at University of Hyderabad.

Several Bilingual and Multilingual dictionaries involving Sanskrit and Indian Languages are available in print form. But these are not available in e-form. Making them available in e-form will further be useful for both the human as well well as machine translation.

3 Strengthening our Sastras

3.1 Sabdabodha

The pada-vakya-pramaṇa sastras viz. vyakaraṇa,mimaṃsaand nyaya sastras deal with the problem of communication through a language. The process of communication starting from expressing the thoughts through a language string by a speaker to understanding a speech by a listener has been well analysed in the texts of Indian grammatical tradition. While these texts and concepts discussed therein have been useful for the analysis of Sanskrit texts, they are also general and hence are useful for the analysis of other languages as well. Some of the important concepts that are found to be useful for the processing of languages are the factors useful for the verbal cognition viz. *akan̲kṣa*, *yogyatạ tatparya*, *sannidhi*, the concepts of *pravṛttinimitta*, *abhidha*, *lakṣaṇạ vyañjana*that deal with various levels of meaning, the concepts of *rḍha artha* and *yaugika-artha* categorising the meanings, the concepts of vṛtti and various types of vṛttis leading to derivation of new words, the concepts of karakas, and the analysis of sentences, integrity of sentences and paragraphs through the concepts of vakaikavakyataand padaikavakyata These concepts have been dealt very well in the tradition discussing several examples from Sanskrit texts, thereby showing its use in practical life in understanding the communication / texts. These theories therefore, can be used a) to develop tools for analysis of Sanskrit texts, and b) to develop Sanskrit related Machine Translation systems. Further, application of these theories to the field of natural Language Processing will naturally throw several challenges which in turn will enrich and strengthen our sastras.

3.1.1 Text processing and Sabdabodha

With the advent of technology, it is now possible to improve the accessibility of the Sanskrit texts which have a wealth of information of various kinds such as scientific, cultural, historic, and so on. There are several ongoing efforts in the field of Sanskrit Computational Linguistics that use the theories of sabdabodha and the Paṇini's Aṣṭadhyay to develop various computational tools. Some of the major efforts have resulted into the analysers through web-services such as http://sanskrit.uohyd.ac.in/scl, http://sanskrit.jnu.ac.in/, http://sanskrit.inria.fr, http://sanskritlibrary.org, http://www.sanskritreader.de, http://www.taralabalu.org/panini, and http://www.sanskritworld.in/sanskrittool/SanskritVerb/tiGanta.html, to name a few.

With these tools, now it is possible

- to generate subantas, tin̲antas, kr̲dantas, taddhitantas, samasta padas, etc.
- to analyse a Sanskrit word

to join two words following sandhi rules

to split a sandhied word

to get the karaka vislesạnạ of a sentence

to search the meaning of a word in various dictionaries

to search a given word in various texts

However, these tools are far from being complete. In addition to improving these tools, we also need

to integrate morph analyser and sandhi splitter into a search engine in order to search the items in a sam̲hita texts.

to develop discourse analysers

to port existing tools to android and similar Operating Systems for wider usability

spell checkers and grammar checkers integrated in the editors

predictive text editors

analysers to handle Vedic texts

analysers for meters (chandas)

reliable Machine translation systems that help users to understand more popular texts such as Geeta, etc.

Grammar tools for Ayurveda students to understand Ayurveda texts

3.1.2 Machine Translation systems

In addition to the development of Sanskrit-Hindi Machine Translation system the concepts in the Indian tradition are also being used to develop Machine Translation system among Indian languages as well as Machine Translation sys-tem from English to Hindi. The claim here is that with the knowledge of a few grammatical concepts useful for the verbal cognition, with the help of a ma-chine translation system, it should be possible to access texts in ANY language through the mother tongue. This will result in not only the strengthening of mother tongue, but also give a boost to the study of Indian grammatical theories with a new perspective. However these efforts are being carried out in isolation. It requires a special attention and support in the form of various resources { both financial as well as human.

3.2 Study of Aṣṭadhyay: New dimensions

The study of Aṣṭadhyay is gaining new dimensions in the era of Information Technology. It is considered to be the first ‘formal programme’ written for human beings, in a well defined formal language derived from the then prevalent Sanskrit language. Computer scientists are looking at the organisation of the sutras, its compactness, the conflict resolution it uses, and the use of Natural Language in writing the sutras. Its importance as a nearly full-fledged grammar is well known, and there are efforts to build computational tools, following this grammar, to analyse Sanskrit language texts. The third importance of this great monument of human intelligence is the universality of the concepts of language analysis. There have been efforts in all these three areas all over the world, though at very minuscule level.

3.3 Logic, Ontology and Discourse Theories

The theories of pramaṇa such as Nyaya, the knowledge representation as discussed in Navya Nyaya, the discourse theories discussing the sangatis, the ontological classification of the entities by Vaisesịkas have all relevance in the processing of Natural Languages. Various efforts in this direction in each of the fields by researchers in the recent past have shown that use of Indian grammatical theories and knowledge systems do yield positive results.

4 Conclusion

In the conclusion, I first point out how we missed the opportunities earlier, and what are the possible threats, if we do not rise now. Next I provide clues to attract and encourage the young scholars to this field and finally mention potential benefits in this area.

4.1 Loosing the battle and a possible threat

The researchers in the West working on Machine Translation systems have observed that the dependency grammars which are close to our traditional Karaka analysis are more suitable for the language analysis. There have been efforts all over the world towards the development of Dependency Analysers for various languages. Had the Sanskrit scholars been involved in the field of NLP research in India, we could have taken a lead in this area. However, hardly any Sanskrit scholar, with an exception of one scholar Prof. K V Ramakrishnamacharyulu, was involved in this work. The picture is not still dark. Still there is a need to develop dependency grammars for all Indian languages. Before the West develops dependency grammars for our languages, let the Sanskrit scholars rise to the occasion, and use their knowledge of Sanskrit grammar and develop dependency grammars for their own Mother tongues so that the Natural Language Processing community gets benefited.

There are other possibilities as well where we can contribute. The problem of Word Sense Disambiguation is one such area. Another area is the area of discourse analysis. What is needed is a will power to deviate a little from the tradition and take a calculated risk of exploring a new path.

At this point I would also like to point out a possible threat. There have been efforts to build Machine Translation systems using statistical and machine learning techniques. It is not far off that such systems be available for all Indian languages as well. If we are serious enough to show the relevance of satras and take advantage of them to build a better system, still there are hopes.

5 Possible ways of nurturing this field

In order to promote the new emerging field of Sanskrit computational linguistics, it needs to be nurtured promoting teaching, research and development, all the three together.

Conduct teacher's training programs introducing the teachers to this emerging inter-disciplinary area

Introduction of inter-disciplinary courses on
a) Sanskrit Computational Linguistics
b) Applied grammar
c) Machine Translation

Promote various Development activities such as
a) Workshop on developing Tagging guidelines
b) Projects for building and improving the computational tools such as OCR, speech recognition and generation, search engines suitable for searching Sanskrit texts
c) Development of Dependency grammars for Modern Indian Languages

Foster Basic Research pertaining to the use of
a)sabdabodha concepts for analysis of texts other then Sanskrit
b)Navya-Nyaya for knowledge representation
c)Discourse theories of Mmm sa for discourse analysis
d)Vaiseṣika Ontology for meaning disambiguation

Provide Incentives to
a) Engineering students to take up projects related to Sanskrit
b) M.Phil. projects in the area of Sanskrit Computational Linguistics

5.1 Potential benefits

Potential benefits of the synergy of Indian systems with the modern technology are many.

- The study of our saatras will get a boost
- The link between Sanskrit and Indian languages will get strengthened.
- The NLP industry is expected to grow to 13 billion dolor by the end of 2020. India being multi-lingual, there are huge growth possibilities in this sector in India as well. This will open up several job opportunities for our students.