An effort to develop a tagged lexical resource for Sanskrit

S. Varakhedi, V. Jaddipal, V. Sheeba

AN EFFORT TO DEVELOP A TAGGED LEXICAL RESOURCE FOR SANSKRIT

S. Varakhedi
V.Jaddipal
V. Sheeba
Rashtriya Sanskrit Vidyapeetha Deemed University
Tirupati
{shrivara,v.jaddipal,v.sheeba}@gmail.com

1. ABSTRACT

In this paper we present our efforts the first time of its kind in the history of Sanskrit to design and develop a structured electronic lexical Resource by tagging a Traditional Sanskrit dictionary. We narrate how the whole unstructured raw text of Vaacaspatyam – an encyclopedic type of Sanskrit Dictionary has been tagged to form a user friendly e-lexicon with structured and segregated information through corpus designing methods.

2. INTRODUCTION

It is not unknown to the scholars in the field of computational linguistics that electronic lexical resources are useful not only for human understanding but also for the needs of language processing. Many NLP applications like Morphological Analyzer etc. inevitably require a well-formed lexical resource. Lexical resource with grammatical and semantic information would be helpful in processing and in translation and decision making inference engines as well. It is not possible to do any kind of language processing in syntactic and semantic level without the structured relevant information regarding the stems of that language. Keeping this in view we have developed this e-resource for the Sanskrit language.

The Sanskrit language is one of the oldest classical languages of the world. It has a gigantic literary treasure related to all branches of sciences and all walks of life. Sanskrit is the first language to have a very precise grammar formalism authored by Paa.nini two thousand years ago. No other language has such a great tradition of grammar formalism, which is sound, perfect and very formal in nature. For these reasons Sanskrit gives enormous scope for NLP researchers and computer scientists from computational view point. With out proper lexical resources, NLP researchers will find themselves handicapped. This is the meeting point of traditional linguists and computational researchers. The information available in conventional lexicons are not sufficient for computational analysis. The way how information is stored becomes more crucial rather than how much information is available in the lexicon. Therefore the lexicographer of an electronic lexicon should be careful while designing the lexicon for computational processing purpose. In case of Sanskrit the design of e-lexicon is more complex because the traditionally available lexicons or dictionaries in Sanskrit have by nature structural complexities. They are designed in an organized but not so well organized for computational purposes. Nevertheless they become important for they carry tremendous and immeasurable information. Hence restructuring of such available dictionaries in Sanskrit is the pre-eminent necessity of Sanskrit computational linguistics. In this direction, our team has opted Vaacaspatyam for tagging on an experiment basis with a goal of developing a multipurpose electronic lexical resource that could be useful for academic research and computational processing as well. This work opens up further a new avenue of research in development of Sanskrit electronic dictionaries in a similar method. The experience gained through this program has prompted us to take up V.S.Apte Sanskrit-English Dictionary in hand.

3. HARD-BOOK TO SOFT-BOOK

This encyclopedic type of lexicon was first published in Kolkata in 1884. It is needless to say that the re-printed editions that are available now are not at all in readable condition due to old font types, unclear print and missing characters. Apart from all this,

untraced errors in the original print often mislead the readers. It was felt necessary to have an electronic version of this gigantic work which runs into about 11 thousand pages in six big volumes printed in old kolkata printing halls using small fonts. There are no breaks in words, not even clear breaks in topics and paragraphs. For each entry tremendous information is collected. Nevertheless, everything is undivided and hence not easily accessible even by eminent scholars. To avoid all this trouble, we initiated to develop an e-content of the Vaacaspatyam, which can be searched and retrieved with out any hustle. We started keying in the data into machines in ISCII using gist technology developed by CDAC Pune. Within a year we got the first raw version of the original text that needed several readings to get the proof corrected.

4. INTRODUCTION TO THE VAACASPATYAM

Pandit **Taaranaatha Tarkavaacaspati**, with his in-depth erudition and indefatigable industriousness devoted several years to prepare encyclopedic Sanskrit Lexicon called 'Vaacaspatyam', consisting of about 5442 printed pages of A4 size. It contains terms along with their derivations and explanations drawn from almost all the branches of Sanskrit Literature, such as the Vedaas, Vedaa"ngaas, Puraa.naas, Upapuraa.naas, Philosophy, Tantra, Artha"saastra, Ala"nkaara"saastra, Chan.da"s"saastra, Sangitasastra, Military Science, Paakavi.dyaa, "Sik.sa, Kalpa, Hasti "Saastram, Ha.tha-Yoga and Vaastu"saastra etc. Besides these, the technical words and doctrines of the following systems of Philosophy are fully explained: Caarvaaka, Maadhyamika, Yogaacaara, Vaibha.sika, Soutraantika, Arhata, Raamaanuja, Maadhva, Paa"supata, "Saiva, Pratyabhij˜na, Rase"swara, Paa.niiya Vyaakara.na, Nyaaya, Vai"se.sika, Mimaamsa, Saa"nkhya, Pata˜njali-Yoga "Sastra and Vedaanta.

It is needless to say that lexicon is an essential part of language -learning. The most ancient language of the world, Sanskrit, also has had Lexicons such as Amarakosa, Vaijayanti, Vi"swaroopako"sa, Naanaarthako"sa etc., which serve more like Thesaurus than dictionary. However, dictionaries such as "Sabdakalpadruma of Raadhaakaantadeva were compiled, where in the grammatical requirements of the Sanskrit reader were also met with. The Vaacaspatyam as a lexicon of Sanskrit words with word-derivation, grammatical specification as per the Paa.ninian System is an excellent work for reference concerning the Sanskrit system of word-meaning. The uniqueness of this lexicon, in comparison to even its succeeding lexicons such as Apte's dictionary, is that while most dictionaries concentrate on the semantic aspect of words listed, the Vaacaspatyam not only deals with their grammatical aspects but also gives all the details available in Sanskrit literature. Though the Vaacaspatyam is constructed in the style of "Sabdakalpadruma, it excels "Sabdakalpadruma in references and size.

This Sanskrit lexicon which is encyclopedic in nature is a pioneering work of its kind and ever since its publication has been held in the highest esteem not only in India but even in England and Europe, because it is by far, wider and deeper in scope than any other contemporary Sanskrit dictionary.

The author of the work Pandit Taaraanatha Tarkavaacaspati himself states that in addition to all the derivations and different meanings with illustrations of all the words which are found in Wilson's Sanskrit Dictionary and Raja Raadhaakanta's "Sabdakalpadruma, Vaacaspatyam contains numerous Vedic words which are not found even in the Bohtlink's St. Petersburg Sanskrit - German dictionary.

5. CONTENTS AND FEATURES OF VACHASPATYAM

1. Listing of words in alphabetical order.
2. Paa.nini's li"ngaanu"saasana on genders.
3. Panini's rules on the suffixes.
4. Paa.nini's rules on the primitive and derivative words.
5. The derivation and different meanings with illustrations of all the words which are found in the Wilson's Sanskrit dictionary and Raadhaakaanta's "Sabdakalpadruma and numerous words not to be found in the said or any previous dictionaries.
6. The derivations and different meanings of the words of the Vedas.
7. Numerous Vedic words not to be found in Bohtlink's Sanskrit and German dictio-

nary.Technical words and doctrines of the following system of Philosophy are fully explained - Caarvaaka, Maadhyamika, Yogacara, Vaibha.sika , Soutraantika, Aarhata, Raamaanuja, Maadhva, Paa"supata, "Saiva, Pratyabhij˜na, Rase"swara, Paa.nini, Nyaaya, Vai"se"sika, Miimaamsa, Saankhya, Paatanjala-Yoga and Vedaanta.

8. The technical terms of the "Srouta and G.rhya sutras.

9. The technical words of Sm.ritis.

10. The plan and scope of all the Puraa.naas and Upapuraa.naas.

11. Plan and scope, of the Mahaabhaarata and Raamaaya.na.

12. The History of the ancient Kings of India as far as gathered from the Puraa.naas and Upapuraa.nas.

13. The position of the different countries/dwiipaas according to ancient Indian Texts. (Brahmaan.davar.nanam)

14. The full explanation of technical terms of Ayurveda and also an account of Ancient Indian Science of Anatomy and the preparation of the medicines according to Ayurvedic texts.

6. DEVELOPMENT OF E-VAACASPATYAM

The Vaacaspatyam contains about 46970 unique word entries with explanation. Each entry has a minimum of 2 lines of information and in most of the cases it runs about 10 to 20 lines. Some prominent words may have elaborate entries upto 20 pages on their category, meaning, sources, usages and other related information. More than 200 source books of different branches and disciplines of learning and are cited. References of more than 30 ko"shaas (lexicons) are found. Names of these references and sources that were cited in short abbreviations are expanded to their full form for the benefit of the readers. The tags help in segregating such flowing information.

7. TAGGING SCHEME FOR E-VAACASPATYAM

As soon as the basic electronic text was ready for application, we started to tag the text with meta-tags to identify the structures of the lexicon. The following tagset was developed for marking.

Tag Description Example

1. <cat> Category of the word a <cat> avyaya </cat>

2. <vp> Vyutpatti i.e., etymology aja <vp> na jaayate </vp>

3. <pr> PrayogaH – usage in any standard Sanskrit work

4. <vkr> Vyakarana information aja<vkr> na+janii+da </vkr>

5. <ar> Artha i.e., meaning aja <ar> caturmuKa </ar>

6. <vg> Vigraha i.e, explanation given for compound word

7. <akr> Aakara i.e., source for the word or its etymology, usage etc.

8 <vn> Vivara.nam i.e., narration about the particular concept

This tag set is used to segregate the semantic part of the text. For stylistic presentation, we have used some other tags (<sl> to indicate "Sloka etc.) which are not listed here. Initially the tagging was done manually with help of some scholars. Later, we could find out some heuristics using which 70% of the text was mechanically tagged. Through this method we saved human labor as well as money.

The raw text added with these meta-tags, which contain necessary information about linguistic features, has become a good resource not only for presentation but also for better understanding of the original text and the semantics of the words listed in the lexicon.

This Tagging scheme has given tree structure to the lexicon in the following way.

A) Simple structure 1
<word> %stem%
<category> INFO </category>
<grammar> INFO </grammar>
<meaning1> INFO </meaning1>
<usage1> INFO </usage1> . . . .
<ref1> INFO </ref1> . . . . . .
<meaning2> INFO </meaning2> . . . . . .
</word>

B) Structure 2
<word> %stem%
<grammar> INFO </grammar>
<category1> INFO </category1>
<meaning1> INFO </meaning1>
<usage1> INFO </usage1> . . . .

<ref1> INFO </ref1>....
<usage2> INFO </usage2>
<meaning2> INFO </meaning2> .....
<category2> INFO </catgory2>
<meaning1> INFO </meaning1>
<usage1> INFO </usage1> .....
<ref1> INFO </ref1> ...
<usage2> INFO </usage2>
<meaning2> INFO </meaning2> .....
</word>

C) Structure 3

<word> %stem%
<category> INFO </category>
<grammar1> INFO </grammar1>
<meaning1> INFO </meaning1>
<usage1> INFO </usage1> .....
<ref1> INFO </ref1>....
<usage2> INFO </usage2>
<meaning2> INFO </meaning2> .....
<grammar2> INFO </grammar2>
<meaning1> INFO </meaning1>
<usage1> INFO </usage1> .....
<ref1> INFO </ref1> ...
<usage2> INFO </usage2>
<meaning2> INFO </meaning2> .....
</word>

Thus the lexicon that was readable only by an intelligent scholar, is made very simple in structure in order to be understood by a novice in Sanskrit. This structure enabled us to develop a searchable e-dictionary with different kinds of searchable options.

8. COMPLEXITIES IN TAGGING

Since the source text was very much unstructured from computational aspects though it was humanly understandable, it was not so easy to tag the information in order to get a tree structure. There was no standard and common sequence of information for all entries. The following are some examples of the complexity of the text.

<word>
<cat>INF</cat>
<m1>,<m2>...<mn>
<ref1><ref2>....<refn>
<ex1><ex2>.....<exn>
</word>

However it sometimes goes as follows

<word>
<cat>INF</cat> <gr>INF</gr>
<m1>INF<ref1>INF</ref1><m1>
<m2>INF<ref2>INF</ref2><gr>INF<gr><eg>INF</eg><m2
<m3>.....<mn>
</word>

Further in some places the text has following structure

<word>
<cat>INF</cat> <gr><m1>
<m2><gr><m3><m4><ref2><ref3>
"source"<m4><gr>...
</word>

Even in grammar information there is no common way for representing the same. For example

<word>
<gr>"root" – "meaning of root" "suffix"</gr>
INF
</word>
<word>
<gr>"root" – "meaning of root" "suffix" "meaning of suffix"</gr>
INF
</word>
<word>
<gr>"root" – "meaning of suffix" "suffix"</gr>
INF
</word>

9. E-VAACASPATYAM ON CD ROM

The first version of Vaacapspatyam CD ROM is ready for public release with 6 kinds of search facilities. All 42,000+ word entries are indexed alphabetically. By selecting any word in the word-index, one can access information related to the selected word. In the second option, the user can enter any string he wants to search, by clicking on soft key board designed for Sanskrit alphabets. If the string is available in key word list, Machine calls for relevant information about the string entered by the user. The third search option helps the user in searching all related words to a particular concept like wordnet. This option is unique as it helps the user in getting all the related words while composing poems etc. The Fourth search option is yet another unique experiment for Sanskrit Manuscript editors. In this option two entry boxes

are given, where the user can specify the starting and ending letters of the word missing in the manuscripts. The machine brings all the possible words that begin and end with the specified letters. This option is found very much useful for the editors, while reading damaged manuscript with missing letters and words.[1] Another search option is also given for the user to search for usages and expressions taken from various texts for a particular string or word. One can even search for all words derived from a root or a word with any particular suffix. In addition to all these, word-game enriches the CD with an added value.

We hope that users of this CD-ROM will find it useful for their research and other applications. We also hope that this becomes a model for tagging of any Sanskrit or other Indian language lexicons.

10. TECHNICAL INFORMATION

The CD-R presentation is developed using Visual Basic. The system tools that are available in VB are used. To avoid problems in font display, the textual output is shown in Netscape browser (Version 4.5) using DV-TT-yogesh font developed for iscii text. This method is a tested one and works in all platforms from windows-98 to windows-XP and gets rid of broken font display problem while presenting the text through the browser. At the development stage Perl was extensively used for tagging manually annotated text and for removing errors.

11. FURTHER SCOPE FOR RESEARCH

It is needless to establish that such a work of developing e-lexicon added with information meta-tags is of high importance in language processing. However, the traditional Sanskrit lexicons are rich in content and poor in organization from computational aspects. They need to be restructured for computational purposes. This work poses several challenges for lexicographic science in computational linguistics. There are dozens of such complicated dictionaries like. "Sabdakalpadruma and V.S.Apte dictionary for Sanskrit-English vise-versa. Our team has started working on these both dictionaries. We hope we will come with good results very soon.

12. ACKNOWLEDGEMENT

The authers are thankful to the Vice Chancellor, RSVidyapeetha,Tirupati, K.V. Ramakrishnamacharyulu, Amba P.Kulkarni, Deeptha, Anilkumar, Administrative Heads of Vidyapeetha and Referees of the paper.

13. REFERENCES

1. Taranatha Tarka Vachaspati, Vachaspatyam, (Reprint) Rashtriya Sanskrit Sansthan, New Dehli, 2000.

2. Descartes and Bunce, Programing the Perl DBI, O'Reilly 2000.

3. Erik T. Ray & Jason McIntosh, XML and perl, O'reilly 2002.

4. Jeffrey E.F. Friedl ,Mastering Regular Expressions, O'Reilly 2002.

14. APPENDIX

Figure 1: Home page of Vacaspatyam

[1] See Appendix – II for Search option images in CD-ROM.

Figure 2: Alphabetical-index

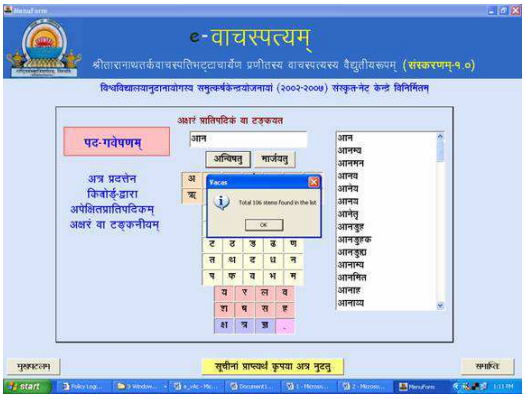

Figure 3: Word-Index

Figure 6: Grammatical Specialities-2

Figure 4: Editor's help

Figure 7: Grammatical Specialities-3