Use of *Amarakośa* and Hindi WordNet in Building a Network of Sanskrit Words

Akshar Bharati, Amba Kulkarni and Sivaja S. Nair

Department of Sanskrit Studies
University of Hyderabad
Hyderabad, India
apksh@uohyd.ernet.in,
sivaja.s.nair@gmail.com

**Abstract**

Sanskrit has a rich source of lexical resources in the form of various kinds of dictionaries, and a thesaurus in the form of *Amarakośa*. Further the rich derivational morphology provides various kinds of relations between the derived words with their head words. With the advent of computational technology now it is possible to build tools that can help a serious reader of Sanskrit to navigate through various words passing through different linkages the word has, to get a holistic view of the meaning of a word, provided such a network exists.

Present work is the first step in that direction. We have initiated the process of building a network of Sanskrit words with *Amarakośa* as the starting point. Since Sanskrit has rich inflectional morphology, we have also linked the web interface to *Amarakośa* with the inflectional morph-analyser. Further to provide various lexical and semantic relations between words, we explored the possibilities of using existing Hindi WordNet. It was found that the comparison of synsets of Hindi WordNet with that of *Amarakośa* is useful in improving the quality of Hindi WordNet on the one hand while enhancing the Sanskrit synsets quantitatively on the other hand.

1 Introduction

Ever since the development of English WordNet (Fellbaum, 1999) the computational lexicography work has gained momentum and acquired a new direction. Several projects purely dedicated to building WordNets for different languages, linking the existing WordNets and building multilingual wordnets were taken up during the last decade (Vossen, 2002 and Sinha et. al, 2006). Though the usefulness of WordNet for NLP is still to be established, there are several efforts to show its significance and relevance for the NLP related work (Agirre E. et. al, 1996).

In India, there have been efforts at several places all over the country to develop WordNets for Indian Languages (Tamil, Marathi, Hindi, Sanskrit) (Tamil WordNet, Marathi WordNet, Hindi WordNet and Sanskrit WordNet). Sanskrit being the mother of several Indian languages, it is natural to think of Sanskrit WordNet at the central place linking all other Indian Languages. Though there were initiatives to start the work on Sanskrit WordNet (Mohanty et. al, 2002) nothing concrete has yet come out.

In the next section, we describe the nature of Sanskrit language, and the available lexical resources. The third section mainly describes the lexical database built from the *Amarakośa* - the oldest lexicographic text on non-vedic Sanskrit. The fourth section discusses the feasibility of building Sanskrit WordNet based on the existing Hindi WordNet, with *amarakośa* as the starting point. We conclude by identifying the tasks that need to be carried out in order to build a usable network of Sanskrit words.
2 Word Formation in Sanskrit

Two important aspects of language study are its grammar and its lexicon. Pāṇini’s *Aṣṭādhyāyi* and Amarasimha’s *Nāmālīṁgamuṣasanaṇam* popularly known as *Amarakośa* both belonging to roughly 5th century B.C. serve as monumental works in the area of grammar and lexicography respectively. Though lexicographic works such as *Nighantu* existed before *Amarakośa*, *Amarakośa* dealt with essentially non-vedic words and hence gained importance very soon.

Some languages build extensively while others to a limited extent only. Raguvira(1981) in the introduction of his ambitious project of building English - Hindi dictionary of technical terms, where he borrows heavily from Sanskrit, describes the richness of word-formation in Sanskrit in the following words.

*While every language builds to a certain extent, it is only a very small number that build constantly, and not only single stray words but whole systems. These are the three great classical languages of the world. ... are Sanskrit, Chinese and Latin (with Greek)(Raghuvira, 1981).*

Figure 1 describes the rich word formation in Sanskrit through the Finite State Transducer(FST).

Thus, as is clear from figure 1, the relation between words across Part of Speech(POS) also becomes very significant in case of Sanskrit. However English WordNet does not contain syntagmatic relations linking words from different syntactic categories except for a few such as legal-lawyer, big-size(Fellbaum, 1999). To get an idea of the richness in building words in Sanskrit, we show in figure 2 the compositional meaning of nouns derived from verbs by adding non-finite suffixes(*kṛt*). Sanskrit has around 140 such *kṛt* suffixes, and the derivation is quite productive. As is evident from the figure 2, such a network of Sanskrit words explaining the relationships among them is a valuable resource for any NLP work related to Sanskrit. The important role of verbs in

---

**Figure 1: Word Formation in Sanskrit**

<table>
<thead>
<tr>
<th>dhāatu</th>
<th>verbal root</th>
<th>sup</th>
<th>nominal suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>subanta</td>
<td>noun</td>
<td>kṛut</td>
<td>nonfinite verbal suffix</td>
</tr>
<tr>
<td>ting</td>
<td>finite verbal suffix</td>
<td>shabdakosha</td>
<td>lexicon</td>
</tr>
<tr>
<td>dhāataupāatha</td>
<td>verbal root</td>
<td>taddhita</td>
<td>derivational suffix</td>
</tr>
<tr>
<td>sanaadi</td>
<td>derivational suffixes</td>
<td>upasarya</td>
<td>verbal prefix</td>
</tr>
</tbody>
</table>

---

**Figure 2: Sample Derivation in Sanskrit**

[kartāi (to do)]

- kartas/kurvas/kaarakā
- karma/karvayakriyamaana
- bhaavi/acti
- kartas/kurvas/kaarakā
- kartas/kurvas/kaarakā
- pakta/pakani/pacaka
- pakta/pakani/pacaka
- pakta/pakani/pacaka
- pakta/pakani/pacaka

[karina (action of cooking)]

- bhaavi/acti
- pakta/pakani/pacaka
- pakta/pakani/pacaka
- pakta/pakani/pacaka
building Sanskrit WordNet is also highlighted by Kulkarni(Kulkarni et. al, 2008).

Thus there are two distinct tasks: one is to
develop a network of words within a syntactic
category which is more or less parallel to
the concept of English WordNet, though in
case of Sanskrit the ontological classification
may be influenced by the Vaiśeṣika ontology.
Owing to the productive nature of Sanskrit in
word building there is another important and
unique task of developing a network between
the words belonging to different syntactic
categories but related semantically. In this
paper, we take a stock of existing resources
and show one can benefit from these to
accomplish the first task, restricting ourselves
to the nouns only.

3 Existing Resources

The tradition of lexicography is very old
in Sanskrit. Sanskrit literature is rich with
many lexical resources such as Nighaṇṭu,
Amarakośa, Vācaspatyam, Śābdakalpadruma,
etc. Sanskrit lexicographical work falls
broadly under two categories: work related
to Vedic Sanskrit and the work related to
the laukika Sanskrit - the language which
is in normal use. Amarakośa is the first
exhaustive lexicographic work falls
broadly under two categories: work related
to Vedic Sanskrit and the work related to
the laukika Sanskrit - the language which
is in normal use. Amarakośa is the first
exhaustive lexicographic work which
has been the source for many
commentaries, and derived works. It has
three chapters(kāṇḍa): the first chapter
mainly deals with the words either related to
pañcamahābhūta(five elements) or abstract
concepts such as dik(direction), kāla(time),
vāk, etc., whereas the second chapter mainly
deals with the actual realities such as human
beings, animals, plants, etc. The third
chapter is essentially a residue with a major
part devoted to polysemous words. Since
the Amarakośa words cover commonly used
words, it is thus natural to start the work
with core words from the Amarakośa.

The other important resource is the existing
Hindi WordNet(Hindi WordNet). Hindi is
basically an offshoot of Sanskrit, though it has
many words of Arabic or Persian origin. The
Hindi WordNet has around 27,879 synsets
and has its own ontology which is different
from that of English WordNet and has around
200 ontological classes as against 25 unique
beginners used in English WordNet.

4 Our Work

Our goal is to build an electronic network
of Sanskrit words, showing various relations
among the words. The relations may be either
lexical or semantic, and may be between
words within the same category or may be
between the words across categories. In this
presentation, we concentrate only on the
relations between words belonging to the
same categories, covering only nouns.

It is natural to base the work on Amarakośa
as it has around 9990 words of which 9036
are distinct. Considering the vocabulary of
Sanskrit, this figure may look very small.
However these are the very frequently used
words in day-to-day life and hence have
special importance.

4.1 Lexical Database of Amarakośa

The text of Amarakośa is in the form of verses
composed mainly in anuṣṭupa meter. These
verses list the synonymous words and also
indicate the gender of the words wherever
necessary. In the beginning of Amarakośa
some default rules for assigning gender to the
words are given. Later wherever necessary the
exceptions are mentioned separately. There
are also certain words solely used for the sake
of completion of meter. Ignoring such words
which indicate the gender and the words
which are used for completion of meter, all
other words have been entered in the database
as shown in table 1.

<table>
<thead>
<tr>
<th>Word</th>
<th>Chapter-Varga-Verse-Line</th>
<th>Gender</th>
<th>Class</th>
<th>Synset-id-word</th>
</tr>
</thead>
<tbody>
<tr>
<td>amara</td>
<td>1.1.7.1</td>
<td>puM.</td>
<td>svargavarga</td>
<td>svarga</td>
</tr>
</tbody>
</table>

Table 1: Sample entry in the database

The synset-id-word is an unique identifier
indicating the synset the word belongs to. All
the words having same synset-id-word forms one synset. For example, table 2 shows a sample synset.

<table>
<thead>
<tr>
<th>Word</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>chada</td>
<td>2.4.14.1</td>
</tr>
<tr>
<td>chadana</td>
<td>2.4.14.1</td>
</tr>
<tr>
<td>paläśa</td>
<td>2.4.14.1</td>
</tr>
<tr>
<td>parṇa</td>
<td>2.4.14.1</td>
</tr>
<tr>
<td>dala</td>
<td>2.4.14.1</td>
</tr>
</tbody>
</table>

Table 2: Sample synset

A polysemous word belongs to more than one synset, as shown below.

**patra**

*chada, chadana, paläśa, parṇa, patra, dala*

synset-id-word = *patram* (leaf) Reference = 2.4.14.1

*chada, garut, tanüruh, pakṣa, patra, patra*

synset-id-word = *pakśipakṣah* (wing) Reference = 2.5.36.1

**patra, vāhana, dhoraṇa, yana, yugya**

synset-id-word = *vāhanam* (vehicle) Reference = 2.8.58.1

A section of third chapter of *amarakośa* contains a list of polysemous words with different meanings. To avoid duplication, only the meanings that have not been covered in earlier chapters have been entered.

The database has 9990 records with 9036 distinct words and 4062 distinct synset-id-words (or Synsets). The table 3 shows number of polysemous words with the polysemy count, with examples for the first few.

<table>
<thead>
<tr>
<th>meanings</th>
<th>words</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>1</td>
<td>hari</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>go, antara</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>puskara</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>kuṭa</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>vṛkṣa, kriyā, aksa</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>suci, rasa, ghana, bala, bhaga</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>dhātu, dharma, etc.</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>179</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>368</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>893</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7458</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Polysemy Distribution

the *Amarakośa* that displays different synsets associated with a given word. A tool-tip displays *Amarakośa* reference of a word along with its gender.

A web based interface (Amarakosha interface) has been developed to display the synsets covering various meanings of the given word, along with the gender information. Taking into account the inflectional richness of the Sanskrit language, the input is filtered through the morphological analyser for possible inflections.

Figure 3 is a snapshot of the interface of the *Amarakośa* that displays different synsets associated with a given word. A tool-tip displays *Amarakośa* reference of a word along with its gender.

Figure 3: Snapshot of the web display

5 Comparison of *Amarakośa* Synsets with Hindi WordNet synsets

A good coverage Hindi WordNet with around 27,879 synsets and around 200 unique beginners is available. Hindi being an offshoot of Sanskrit, naturally shares a lot with Sanskrit both at the syntactic as well as semantic level. It is natural therefore to expect that a large part of the synsets will be common to both Sanskrit and Hindi. An experiment was carried out to measure the overlap between the synsets from *Amarakośa* and those from Hindi WordNet. *Amarakośa* has 4062 synsets
whereas Hindi WordNet has 27,879 synsets. Among these, only 1782 concepts ‘matched’. Though the match was perfect at the conceptual level, there are some observations:

- Hindi WordNet has some synsets whose entries need to be corrected. For example, the word *śambhu* has been entered in two synsets

  Synset ID: 00002061
  Synset: *śiva:* *śamkara:...:* *śambhu:...*  
  Concept: *eka *śrītināśaka *hindu devatā*  
  gloss: Hindu god who is destroyer of the universe.

  Synset ID: 00002198
  Synset: *brahmā:* *caturānana: *pitāmaha:*  
  *brahmadeva: vidhātā:*  
  *paṇikajātana:* *śambhu: girāpati:* ...  
  Concept: *hinduoM *ke *eka *devatā *jo *śrīti *ke *srjaka*māne *jāte haiM*  
  gloss: Hindu god who is creator of the universe.

As one can see the two concepts are contradictory. *Amarakośa* lists *śambhu* only in the synset corresponding to the first concept where it should be.

- In several cases there is a fine-grain distinction. For example, the words such as *haridrā* or *pālāsa* may stand for both the tree as well as its fruit. Hindi WordNet distinguishes between these two concepts, whereas *Amarakośa* does not.

Present work is the first step in that direction. We have initiated the process of building a network of Sanskrit words with *Amarakośa* as the starting point. Since Sanskrit has rich inflectional morphology, we have also linked the web interface to *Amarakośa* with the inflectional morph-analyser. Further to provide various lexical and semantic relations between words, we explored the possibilities of using the existing Hindi WordNet. Since the Sanskrit literature uses *Vaiśeṣika* ontology, the work on comparing the ontology used by Hindi WordNet with that of *Vaiśeṣika* ontology is in progress.

The comparison of synsets of Hindi WordNet with that of *Amarakośa* is useful in improving the quality of Hindi WordNet on the one hand while enhancing the Sanskrit synsets quantitatively on the other hand.

Finally taking into account the Sanskrit’s unique power of building whole system of words, it is utmost important to provide a facility to build a network of words across the POS categories which is absent in the design of WordNet.

7 Acknowledgment

Authors thank K V RamKrishnamacharyulu for useful discussions at various stages of the work.

6 Conclusion

Sanskrit has a rich source of lexical resources in the form of various kinds of dictionaries, and a thesaurus in the form of *Amarakośa*. Further the rich derivational morphology provides various kinds of relations between the derived words with their head words. With the advent of computational technology now it is possible to build tools that can help a serious reader of Sanskrit to navigate through various words passing through different linkages the word has, so that he gets a holistic view of the meaning of a word, provided such a network exists.

References


*Amarakosha* Interface
http://sanskrit.uohyd.ernet.in/~anusaaraka/sanskrit/amarakosha/.


Hindi wordNet: http://www.cfilt.iitb.ac.in/wordnet/webhwn

Marathi WordNet: http://www.cfilt.iitb.ac.in/wordnet/webmwn


