Introducing Sanskrit Wordnet

Malhar Kulkarni
Department of Humanities and Social Sciences,
Indian Institute of Technology Bombay
malhar@iitb.ac.in

Chaitali Dangarikar
Center for Indian Language Technology,
Indian Institute of Technology Bombay
chaitali.dangarikar@gmail.com

Irawati Kulkarni
Center for Indian Language Technology,
Indian Institute of Technology Bombay
irawatikulkarni@gmail.com

Abhishek Nanda
Center for Indian Language Technology,
Indian Institute of Technology Bombay
abhi.nanda@gmail.com

Pushpak Bhattacharyya
Center for Indian Language Technology,
Indian Institute of Technology Bombay
pb@cse.iitb.ac.in

Abstract

How does one build the wordnet of a language that has a rich lexical tradition spanning over millennia? The sheer volume of words and their nuances, the rich, deep and diverse grammatical tradition, the pressure of modern developments on the language- all these factors and more combine to pose unique challenges in creating lexical resources for such languages. This present paper describes the construction of Sanskrit wordnet, being built using the *expansion approach*. It presents the processes and challenges involved in this task that purports to uncover the intimate linkage that underlies Indian languages most of which have speaker population numbering 20 to 500 million.

1 Introduction

Sanskrit is historically an Indo-Aryan language (Deshpande, 1992) and one of the 22 official languages of India. It has a vast literature and the interest in analyzing and translating these texts is always on the rise, worldwide.

Specifically, our motivation for building Sanskrit wordnet arises from the following facts:

1. For all languages in the Indo European family in India, the roots can be traced to Sanskrit. A large part of the vocabulary of these languages is derived from Sanskrit which can, therefore, provide the pivot resource for many Indian languages. The speaker population for these languages range from 10 million (Konkani) to 500 million (Hindi/Urdu).

2. Being a heritage language, there is need to digitize and preserve ancient texts in Sanskrit. This activity is greatly helped by word lists. An Optical Character Recognition Device (OCR) for Sanskrit, for example, would need spell correction after scan, and this would need an exhaustive lexicon.

3. Simlarly, there exists real need for translating ancient texts to preserve traditional culture and knowledge. An online wordnet would no doubt be a great help to a translator.

4. Machine aided translation (MAT) is maturing fast, and automatic translation of Sanskrit text is a challenging problem needing wordnet.

5. There is an enormous amount of Sanskrit text which should be available in keyword based searchable form. Text search is greatly helped by wordnets.

6. The tradition of developing lexical resource is very old in Sanskrit. There are diverse *koshas* (traditional and rich monolingual dictionaries) in Sanskrit (see section 1.2 below). Sanskrit wordnet will serve as the **single reference** point representing and pointing to all these resources.

1.1 Sanskrit language

Indian subcontinent is inhabited by a very large population who speak languages belonging to 4 major families, Indo-Aryan (a subfamily of Indo-European), Dravidian, Tibeto-Burman and Austro-Asiatic. Sanskrit is the oldest member of the Indo-Aryan language

family, a sub branch of Indo-Iranian, which in turn is a branch of Indo European language family.

There is a traditional fourfold division of lexical units of Indian languages into:

1. तत्सम *tatsama*[1]- words having their origin in Sanskrit and accepted in the modern Indo-Aryan languages without any change in their phonology.

2. तद्भव *tadbhava*[2]- words which have their origin in Sanskrit but their phonological forms are changed as per the rules of the modern Indo-Aryan languages.

3. देशी *desh•* –words which are the native words of the particular language and

4. विदेशी *videsh•* – words borrowed from foreign languages.

The links to तत्सम *tatsama* and तद्भव *tadbhava* words, in particular, will be a great pan-Indian linguistic resource for computational purposes. Table 1 below lists some examples of Sanskrit words in Hindi wordnet[3].

HWN Synset	Tatsam word	HWN synset	English meaning
{तुलसी, पावनी, बहुमंजरी, वृंदा, वृन्दा, वैष्णवी, भारवी, मंजरीक, विश्वपावन, विश्व-पूजिता, पुष्पसारा, त्रिदशमंजरी, त्रिदशमञ्जरी, तीव्रा, पत्रपुष्पा, श्रीमंजरी,श्रीमञ्जरी, अमृता}	तुलसी	तुलसी	basil
	वृन्दा	वृन्दा	
	वैष्णवी	वैष्णवी	
	पावनी	पावनी	
	पत्रपुष्पा	पत्रपुष्पा	
{भौंह,भौं,भ्रू,भृकुटी, तेवर,कोदंड,कोडंड,अबरू}	भ्रू	भ्रू	eyebrow, brow, supercilium
	भृकुटी	भृकुटी	
{पेशी,माँस-पेशी,मांस-पेशी,मॉसपेशी,मांसपेशी,माँस पेशी,मांस पेशी,नस}	पेशी		muscle, musculus
	मांसपेशी		
बैंगन,बैगन,भंटा,भाँटा, शाकबिल्व,शाकबिल्वक, वृंताक,वृन्ताक,नीलवृषा, शाकश्रेष्ठा,वृंताकी, वागुण,वरा, चित्रफला,रक्तकंठ, रक्तकण्ठ,निद्रालु, नीलफला,नटपत्रिका	शाकबिल्व	बैंगन	eggplant, aubergine, mad_apple
	शाकश्रेष्ठा	बैंगन	
	चित्रफला	बैंगन	
	वृन्ताक	बैंगन	
	निद्रालु	बैंगन	
	नीलफल	बैंगन	

Table 1: Tatsama words in the HWN

These representative examples show that the synsets in Hindi wordnet contain 60-70% tatsama (directly borrowed from Sanskrit) words.

[1] *Tatsama Shabda Kosha* (*Tatsama* words dictionary) is published by Kendriya Hindi Nideshalaya, Shiksha Vibhaga, Manava Samsadhana Vikasa Mantralaya, Bharata Sarakara in 1988.

[2] See *Hindi ki Tadbhava Shabdavali* (Sarma, 1968).

[3] www.cfilt.iitb.ac.in/wordnet/webhwn.

1.2 Rich lexical tradition of Sanskrit

Sanskrit has a rich tradition of creating léxica (Kulkarni, 2008). *Nighantu*[4] (700BC) on which Yaska is believed to have written a commentary called *Nirukta* is the oldest known treatise that arranged lexical material from the point of view of *synonymy* as well as *homonymy*, and this tradition continued to *Pali*[5] tradition as well. The first and the foremost popular name of lexicon work in classical Sanskrit is Amarasimha's ***Amarakosha*** (6th century AD) (Oka, 1913). The Catalogous Catalogorum lists at least 40 commentaries on *Amarkosha* alone, which shows how important and popular this synonyms dictionary in ancient India was.

There were many other léxica created more or less in the style of *Amarakosha* which are given in Appendix A (11 of them).

The first modern-day dictionary of Sanskrit was the Sanskrit-English Dictionary compiled by Professor H.H. Wilson and published in 1819 (Wilson, 1819)Two Indian dictionaries came out soon after, namely, the *Shabdakalpadruma*[6] (Deb, 1988) of Pt. Sir Raja Radhakanta Dev and *Vacasptyam*[7] (Bhattacharya, 2003) compiled by Pt Taranatha Tarkavacaspati.

So far the electronic lexical resources available for Sanskrit are mainly online dictionaries.[8] The linguistic resources like *Shabdakalpadruma*

[4] *Nighantu* is Sanskrit term for the collection of words, grouped thematic categories with brief annotations

[5] Pali is a Middle Indo-Aryan language (or Prakrit) of India. It is best known as the language of the earliest extant Buddhist scriptures.

[6] *Shabdakalpadruma* is a first Sanskrit uni-lingual dictionary arranged in the modern alphabetical principles. It gives full quotations and definitions from the original *Koshas* which were unavailable in print at that time. Sets of synonymous words from the traditional *Koshas* are arranged under the headword, followed by the brief gloss. Each entry in the lexicon includes headword, its category, meaning, usages in the Sanskrit texts.

[7] *Vacasptyam* is a modern mono-lingual Sanskrit lexicon. It arranges words in the Sanskrit alphabetical order and gives grammatical information with word derivations as per the traditional Sanskrit grammar. It contains about 46970 unique words. Each entry in the lexicon includes headword, its category, meaning, set of synonymous words, usages and some other information.

[8] The online dictionaries available for Sanskrit are-(1) Monier Williams dictionary < http://webapps.uni-koeln.de/tamil/>, (2) Apte's Sanskrit-English Dictionary < http://www.aa.tufs.ac.jp/~tjun/sktdic/>, (3) Apte's English-Sanskrit Dictionary < http://www.sanskrit-lexicon.uni-koeln.de/aequery/index.html> and (4) Spoken Sanskrit Dictionary: an online hypertext dictionary for Sanskrit - English and English - Sanskrit.< http://spokensanskrit.de/>. Apart from that various scanned versions of the printed dictionaries prepared by European scholars are available at < http://www.sanskrit-lexicon.uni-koeln.de/>.

and *Vaacaspatyam* are vast. For example, a comparison of the entries for the word ***war*** in these electronic dictionaries with the synsets of the same word in the Sanskrit Wordnet is a good indicator of the richness of this lexical tradition in Sanskrit.

1. Spoken Sanskrit Dictionary: (7 words) युद्ध, युध्, संग्राम, समर, आयोधन, आहव , रण्य .
2. Apate's Sanskrit-English Dictionary: (7 words) विग्रहः, संग्रहारः, वैरारंभः, वैरं, संग्रामः, युद्धं, रणं
3. Monier Williams Dictionary: (56 words) अनीक ,अभ्यामर्द ,अम्बरीष ,अरर ,आजि ,आनर्त, आयोधन, आहव, आहाव, कण्ठाल, कन्दल, खज, न, नदनु, निग्रहण , पुष्कर ,प्रविदारण ,प्रसर ,बलज ,भण्डन ,भर ,भीमर ,युत्कार , युद्ध ,योध ,योधन ,रण्य ,राटि ,रु ,वराक ,विदथ ,विदार , विदारण ,विमर्द ,विमर्दन ,शम्बर ,शिलीमुख ,संयत्, संयुग , संस्फोट ,संस्फेट ,संक्रन्द ,संक्रन्दन ,संख्य ,संगथ ,समनीक , समाघात ,समुदय ,समुदाय ,समुद्धर्ष ,समोह ,समर ,समृति , सम्पराय ,हान्त्र ,हार and
4. Sanskrit Wordnet: (97 words) युद्धम् , संग्रामः , समरः, रणः, समरम्, आयोधनम्, आहवम्, रण्यम्, अनीकः, अनीकम्, अभिसम्पातः, अभ्यामर्दः, अररः, आक्रन्दः, योधनम्, जम्यम्, प्रधनम्, प्रविदारणम्, मृधम्, आस्कन्दनम्, संख्यम्, समीकम्, साम्यरायिकम्, कलहः, विग्रहः, संप्रहारः, कलिः, संस्फोटः, संयुगः, समाघातः, अभ्यागमः, आहवः, समुदायः, संयत्, समितिः, आजिः, समित्, युत्, संरावः, आनाहः, सम्परायकः, विदारः, दारणम्, संवित्, सम्परायः, बलजम्, आनर्त्तः, अभिमरः, समुदयः, विवाक्, विखादः, नदनुः, भरः, आक्रन्दः, पृतनाज्यम्, अभीकम्, समीकम्, ममसत्यम्, नेमधिता, सङ्काः, समनम्, मीळ्_हे, पृतनाः, स्पृत्, स्पृद्, मृत्, मृद्, पृत्, पृद्, समत्, समर्यः, समरणम्, समोहः, समिथः, सङ्खः, सङ्गः, संयुगम्, सङ्गथः, सङ्गमः, वृत्रतूर्यम्, पृक्षः, आणिः, शीरसातिः, वाजसातिः, समनीकम्, खलः, खजः, पौंस्यम्, महाधनः, वाजः, अजम्, सद्म, संयत्, संयद्, संवतः

1.3 The process of building the Sanskrit wordnet

There are two methods to develop a Wordnet: (1) Expand method and (2) Merge method (Vossen, 2002). In the first method, a wordnet is constructed based on an existing wordnet. In the second method, sub-Wordnets for specific domains are built and later merged. For Sanskrit Wordnet, the Hindi wordnet is considered as the source resource. Though *expanded* from Hindi wordnet, care was taken to ensure that Sanskrit wordnet captures the real lexical structure of Sanskrit language.

1.4 Expansion approach for Indian language wordnets

Wordnet construction activities in India started in 2000 and the Hindi wordnet[9] (Narayan *et al.,* 2002) is the first one which got released on the Web in 2006. It was built *ab initio* using words from available lexical resources of Hindi. The design of the Hindi wordnet follows the famous English WordNet[10].

While following the expand method, the Sanskrit wordnet follows the hierarchy preservation principle (HPP) (Tufis *et al.,* 2008). In the hierarchy of the Hindi wordnet, if synset H_2 is a hyponym of synset H_1, and the translation equivalents in the Sanskrit wordnet for H_1 and H_2 are S_1 and S_2 respectively, then in the hierarchy of Sanskrit wordnet S_2 should be a hyponym of synset S_1. Thus, in the expansion approach lexicographers are spared the task of establishing afresh semantic relations for the synsets of Sanskrit wordnet. Appendix 2 describes and shows the screenshots of lexicographers' interface for creating the Sanskrit wordnet.

1.5 Synset creation in Sanskrit wordnet

Domains: Initially the Sanskrit wordnet started creating synsets with random synsets from the Hindi Wordnet. Later on, lists of important Sanskrit words were acquired from different sources. University of Hyderabad provided a list of most frequent words in their Sanskrit corpus. It consisted of 8338 words. Another word list available on the indology forum[11] contains a list of 127796 unique words from two major epics of Sanskrit literature: *Ramayana*[12] and *Mahabharata*.[13] The third list is prepared based on the lexicon called *Bharatiya Vyavahara Kosha* (Naravane, 1961). Table 2 shows the part of speech distribution of Naravane's lexicon. It contains 2766 words which are used for 1969 concepts related to the day to day life. Table 3 shows a comparison between the lists of Sanskrit words gleaned from various sources mentioned above.

---

[9] www.cfilt.iitb.ac.in/wordnet/webhwn
[10] Wordnet.princetoon.edu
[11] <http://indology.info>
[12] *Ramayana* is an ancient Sanskrit epic. The Valmiki *Ramayana* is published in 7 volumes, *Baroda*: University of *Baroda Oriental* Institute, 1960-1975.
[13] Mahabharata is one of the two important epics of India. *The Critical Edition of the Mahabharata* is prepared by the Bhandarkar Oriental Institute, Pune from April 1919 to September 1966. It has 19 volumes consisting18 Parvan-s; 89000+ verses in the Constituted Text, and an elaborate Critical Apparatus.

The above mentioned words are organized into **52 domains**.[14] Omitting function words, a core set of concepts was prepared and then by Sept. 2009 synsets for all these core concepts were created.[15]

Nouns	Verbs	Adjectives	Adverbs
1512	225	180	52

Table 2: POS distribution of the synsets created (core concepts)

Sanskrit List 1	**Sanskrit List 2**	**Sanskrit List 3**	**Hindi List 1**
Univ. of Hyderabad most frequent words in Sanskrit (Amba Kulkarni)	Sanskrit Word list (Based on Ramayana and Mahabharata)	Number of Sanskrit Words in Naravane's Bhasha Vyavahar Kosh	Hindi wordnet Total number of unique words
8338	**127796**	**2766**	**105157**

Table 3: Sanskrit word list

While creating synsets the following considerations are kept in mind:

Inserting concepts or glosses in the Sanskrit wordnet: A combination of the glosses given in dictionaries like *Shabdakalpadruma* and the translation of the gloss of the Hindi wordnet synset is used to create the Sanskrit synset glosses. While writing the gloss, complicated सन्धिs *sandhis*[16] and समासs *samAsas* (compounds) are avoided. Whenever lengthy compounds (having 5-6 members) became necessary, the members of the compounds were invariably joined with the hyphen symbol (-) as in: ''अन्य-स्थान-संयोगानुकूल-व्यापार meaning *the activity that is helpful in reaching a place*'' *anya-sthAna-saMyogAnukUla-vyApAraH* where the members of the compounds are अन्य (*anya*) ,स्थान (*sthAna*) ,संयोग (*saMyoga*), अनुकूल (*anukUla*) ,व्यापार (*vyApAra*)[17]. and they are indicated by inserting hyphen. For example- the gloss of a verb in Sanskrit is generally created using technical terms like व्यापार *vyApAra* 'action', जन्य *janya* 'produced,' अनुकूल *anukUla* 'helpful,' etc.[18]

2 Problems faced in the expansion approach

In this section we enumerate the challenges faced in creating the synsets of Sanskrit wordnet in consonance with those of Hindi.

[14] These domains are: 1) Grains and Cereals, 2) Limbs of Humans, 3) Medical treatment, 4) Tools & implements, 5) Worms & Insects, 6) Minerals, 7) Food and Drinks, 8)Games & sports, 9) Ornaments & Trinkets, 10) Household articles, 11) Limbs of animals, 12) Post office, 13) Vegetables, 14) Directions, 15) Country, 16) Religion, 17) Court, 18) Birds, 19) Trees & plants, 20) Dress, 21) Nature, 22) Animals, 23) Fruits, 24) Flowers, 25) Young-ones of animals, 26) Amusement, 27) Spices, 28) Weights & measures, 29) Colours, 30) Relatives, 31) Diseases, 32) Reptiles, 33) Conveyances, 34) Occupations, 35) Education, 36) Time, 37) Government, 38) Verbs, 39) Adverbs, 40) Abstract nouns, 41) Adjectives, 42) Prepositions, 43) Numerals, 44) Conjunctions, 45) Collective words, 46) Pronouns, 47) Ordinals, 48) Feminines, 49) Interjections, 50) War, 51) House, and 52) Miscellaneous.

[15] From this time Sanskrit Wordnet became a part of Indo-WordNet activity which provided a common platform for the lexicographers working on various Indian language Wordnets.

[16] Phonological conjoining

[17] This way of giving definitions is typical of Sanskritic tradition which used to strongly emphasise precision. The long compound simply defines the act of *going*.

[18] So using these expressions, Hindi Wordnet gloss is adapted in following ways- (1){रोना,रुदन करना,आँसू बहाना,क्रंदन करना *ronA, rudana karanA, AMsu bhAnA, krandana karanA*} HWN आँख से आँसू गिराना *AMkha se AMsu girAnA*→ SWN सुख-दुःखयोः भावनावेगात् नेत्राभ्याम् अश्रुपतन-रूपः व्यापारः। *sukha-duHkhayoH bhAvanAvegAt netrAbhyAm aZrupatan-rUpaH vyApAraH*, (2){मारना,पीटना,प्रहार करना, ठोंकना,ठोकना,पिटाई करना,धुनना,धुनाई करना, ताड़ना,प्रताड़ना,रसीद करना *mAranA, pITanA, prahAra karanA, ThokanA, piTAI karanA, dhunanA, dhunAI karanA, tADanA, pratADanA, rasIda karanA* } HWN किसी पर किसी वस्तु आदि से आघात करना *kisi par kisI vastu Adi se AghAta karanA* → SWN कस्मिन् अपि केन अपि वस्तुना आहनन-पूर्वकः व्यापारः। *kasmin api kena api vastunA Ahanana-pUrvakaH vyApAraH* (3) {खरीदना,क्रय करना,मोल लेना,लेना *kharIdanA, kraya karanA, mola lenA, lenA*} HWN पैसे आदि देकर किसी दुकान,व्यक्ति आदि से कुछ सौदा मोल लेना *paise Adi dekar kisI dukAna, vyakti Adi se kuch saudA mol lenA*→ SWN आपणे वस्तु तथा च तन्मूल्यम् एतयोः आदान-प्रदानात्मकः व्यापारः। *ApaNe vastu tathA cha tanmUlyam etayoH AdAna-pradAnAtmakaH vyApAraH,* (4) {रूठना, रूष्ट होना, अनखना, रूसना, रिसाना, फूलना, अनसाना, अनखाना, अनैसना *rUThanA, ruSTa honA, anakhanA, rUsanA, risAnA, phUlanA, anasAnA, anakhAnA*} HWN अप्रसन्न होकर उदासीन,चुप या अलग हो जाना aprasanna hokara udAsIna, cupa yA alaga ho jAnA→ SWN अप्रसन्नताहेतुजन्यः वियोगरूपः औदासीन्यफलजनकः वा व्यापारः। *aprasannatAhetujanyaH viyogarUpaH audAsInyaphalajanakaH vA vyApAraH* (5) {आना:1, पहुँचना, पहुंचना, पधारना, अवना, आगमना *AnaA, pahuMcanA, pahucanA, padhAranA, avanA, AgamanA}* HWN एक स्थान से आकर दूसरे स्थान पर उपस्थित होना *eka stAna se Akara dUsare stAna para upasthita honA* → SWN अन्य-स्थान-वियोग-पूर्वकः अन्य-स्थान-संयोगानुकूल-व्यापारः। *anya-sthAna-viyoga-pUrvakaH anya-sthAna saMyogAnukUla-vyApAraH.*

Difficulty of finding equivalent words:
Sometimes it is difficult to find a Sanskrit equivalent for a Hindi word. For example; the word {चाय} *cAya* (*tea*) is very widely used. The concept of *tea* is explained as follows in the Hindi wordnet:

(1) चाय के पौधे की पत्तियों को पानी में डालकर चीनी, दूध आदि मिलाकर बनाया हुआ पेय पदार्थ
cAya ke paudhe kI pattiyon ko pAnI mein DAlkar cinI dUdha Adi milAkar banAyA huA peya padArtha
(A drink prepared by mixing the leaves of the Tea-plant with sugar, milk and water)

But Sanskrit does not have a word of its own for this concept. Monier Williams in his Sanskrit-English dictionary (MW hereafter) suggests that "चहा" *cahA* (which is actually a Marathi word) should be used as a borrowed word. In the dictionary of spoken Sanskrit we find two different regional words "चाय" *cAya* and "चाया" *cAyA* belonging to the North and South regions of India. The gloss field in the synset of {कषायपेयम्, चायः, चाया, चहा} {*kaSAyapeyaM, cAyaH, cAyA, cahA*} in the Sanskrit wordnet is modified as follows:

(2) चायः चहा एवंविधैः शब्दैः भारतीय-भाषासु प्रसिद्धस्य क्षुपस्य शुष्कपर्णानां चूर्णम् उष्णजले अभिपच्य तस्मिन् द्रवे शर्करादुग्धादीन् संमिश्र्य निर्मितम् उष्णपेयम्।
cAyaH cahA evaMvidhaiH shabdaiH bhAratIya-bhASAsu prasiddhasya kSupasya shuSka-parNAnAM cUrNam uSNajale abhipacya tasmin drave sharkarA-dugdhAdIn saMmishrya nirmitam uSNapeyam
(A hot drink which is prepared by first mixing the leaves of the a plant, which is famous by the names like चहा *cahA*, चाय *cAya*, etc. in the Indian languages, into hot water and then mixing it with sugar and milk)

This change is needed to translate the simple Hindi wordnet gloss. Similarly, for the tree plant, the Hindi wordnet gloss is:

(3) एक पौधा जिसकी पत्तियाँ उबलते हुए पानी में डालकर एक पेय बनाते हैं।
eka paudhA jisakI pattiyAn ubalate hue pAnI mein DAlakar eka peya banAte hein
(A plant- dry leaves of which are boiled in the hot water and a drink is prepared)

and this gloss was modified in SWN as:

(4) चायः चहा एवंविधैः शब्दैः भारतीय-भाषासु प्रसिद्धः क्षुपः- यस्य शुष्क-पर्णानां चूर्णं उष्णजले अभिपच्य तस्मिन् द्रवे शर्करा-दुग्धादीन् संमिश्र्य उष्णपेयं निर्मीयते।
cAyaH cahA evaMvidhaH shabdaiH bhAratIya-bhASAsu prasiddhaH kSupaH yasya shuSk-parNAnAm cUrNaM uSNajals Abhipacya tasmin drave sharkarA-dugdhAdIn saMmIshrya uSNapeyaM nirmIyate
(A plant, which is famous by the names like चहा, चाय, etc. in the Indian languages- dry leaves of which are boiled in the hot water and a drink is prepared)

Difficulties with *examples*:
Generally, examples associated with Hindi synsets are translated only if they *read* sensible when translated into Sanskrit. In some cases, quotations from the Sanskrit texts are included in the example field. A special field has been created to record the source of the quotations. This citation field is incorporated in the lexicographer's interface:

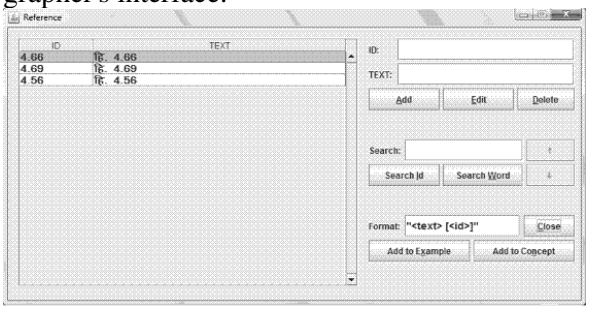

Figure 1. Lexicographer's interface to record citations

The example with the citation is inserted in this format:

(5) "शशि-दिवाकरयोर्ग्रहपीडनम्।"
[भर्तृ.2.91][19]

[19] Here, [भर्तृ.2.91] indicates the place of the quotation in the original Sanskrit text authored by Bhartrhari.

shashi-divAkarayor grahapIDanaM [bhartR 2.91]
(the eclipse of Sun and Moon).

Sometimes, apart from the translation of Hindi example sentence, an alternative example from the Sanskrit text is provided. Multiple examples are separated with the "/" symbol. The sources of the examples are indicated in square brackets. In some cases the translation of the Hindi example sentence becomes problematic due to the unnaturalness of the sentence in Sanskrit.

Coverage of words in Sanskrit wordnet: Taking into consideration the linguistic change and time, it is possible to classify Sanskrit language into three periods- (1) Vedic period-beginning of Vedic Sanskrit can be traced as early as around 1500 BCE and Vedas are written using literals of that time, (2) Classical Sanskrit- A significant form of post-Vedic Sanskrit is found in the Sanskrit beginning with the Hindu Epics—the Ramayana and the Mahabharata. (3) Modern Sanskrit. The usage of the words changed during these periods. The general policy adopted for synset making is to start with the most frequent words of modern Sanskrit and to close the synset with the least frequent word of Vedic Sanskrit. The example of the synset of युद्ध (yuddha) *war*- shown below- is an illustrative case in point. The words in the synset of *war* are arranged from the most common modern Sanskrit words to least used Vedic Sanskrit words.

{युद्धम् , संग्रामः , समरः, रणः, समरम्, आयोधनम्, आहवम्, रण्यम्, अनीकः, अनीकम्, अभिसम्पातः, अभ्यामर्दः, अररः, आक्रन्दः, योधनम्, जम्यम्, प्रधनम्, प्रविदारणम्, मृधम्, आस्कन्दनम्, संख्यम्, समीकम्, साम्परायिकम्, कलहः, विग्रहः, संप्रहारः, कलिः, संस्फोटः, संयुगः, समाघातः, अभ्यागमः, आहवः, समुदायः, संयत्, समितिः, आजिः, समित्, युत्, संरावः, आनाहः, सम्परायकः, विदारः, दारणम्, संवित्, सम्परायः, बलजम्, आनर्त्तः, अभिमरः, समुदयः, विवाक्, विखादः, नदनुः, भरः, आक्रन्दः, पृतनाज्यम्, अभीकम्, समीकम्, ममसत्यम्, नेमधिता, सङ्काः, समनम्, मीळ्_हे, पृतनाः, स्पृत्, स्पृद्, मृत्, मृद्, पृत्, पृद्, समत्, समर्यः, समरणम्, समोहः, समिथः, सङ्खः, सङ्गः, संयुगम्, सङ्गथः, सङ्गमः, वृत्रतूर्यम्, पृक्षः, आणिः, शीरसातिः, वाजसातिः, समनीकम्, खलः, खजः, पौंस्यम्, महाधनः, वाजः, अजम्, सद्म, संयत्, संयद्, संवतः} {*yuddham, saMgramaH, raNaH, samaraH, samaram, Ayodhanam, Ahavam, raNyam, anIkaH, anIkam, abhisampAtaH, abhyAmardaH, araraH, AkrandaH, yodhanam, jamyam, pradhanam, pravidAraNam, mRdham, Askandanam, saMkhyam, samIkam, sAmyarAyikam, kalahaH, vigrahaH, saMprahAraH, kaliH, saMsphoTaH, saMyugaH, samAghAtaH, abhyAgamaH, AhavaH, samudAyaH, saMyat, samitiH, AjiH, samit, yut, saMrAvaH, AnAhaH, saMparAyakaH, vidAraH, dAraNacd, saMvit, saMparAyaH, balajam, AnartaH, abhimAraH, samudayaH, vivAkd, vikhAdaH, nadanuH, bharaH, AkrandaH, pRtanAjyam, abhIkam, samIkam, mamasatyam, nemadhitA, saGkAH, samanam, mIL_he, pRtanAH, spRt, spRd, mRt, mRd, pRt, pRd, samat, samaryaH, samaraNam, samohaH, samithaH, saGkhaH, saGgaH, saMyugacd, saGgathaH, saGgamaH, vRtratUryam. pRkSaH, ANiH, ZIrsAtiH, vAjasAtiH, samanIkam, khalaH, khajaH, pauMsyam, mahAdhanaH, vAjaH, ajaM, sadma, saMyat, saMyad, saMvataH* }

The problem of meaning attestation: Sanskrit has a rich tradition of lexical resources. But the downside of this fact is that the lexicographer has to verify the consistency of word definitions at every step from multiple sources. For example, following words are mentioned in *Shabdakalpadruma*, but other dictionaries prepared by modern scholars like Monier Williams (MW) make the following remarks in the gloss of these words. All of them are used in the Vedic literature for the concept of "war".

सङ्खे	MW- सङ्ख -not found in MW and शब्दकल्पद्रुम
सङ्गे	MW- सङ्ग is not found in sense of युद्ध in MW and शब्दकल्पद्रुम
संयुगे	MW- संयुग n. conflict, battle, war MBh. Ka1v. &c (cf. Naigh. ii , 17)
सङ्गथे	MW- सङ्गथ m. conflict, war Naigh.
सङ्गमे	MW- सङ्गम does not have the sense of युद्ध
वृत्रतूर्ये	MW- वृत्रतूर्य n. conquest of enemies or वृत्र , battle , victory RV.
पृक्ष	MW- पृक्ष m. = संग्राम Naigh. ii , 57.
आणौ	MW- आणि m. (cf. अणि (the pin of the axle of a cart RV. i , 35 , 6 ; 63 , 3 (["battle " Naigh. ii , 17]) and v , 43 , 8
शीरसातौ	शीरसाति is not found in MW and शब्दकल्पद्रुम
समनीके	MW- समनीक n. battle, war RV. (Naigh. ii , 17) Ba1lar. Vii , 60÷61.
खले	MW- खल m. contest, battle Naigh. Nir.
खजे	MW- खज m. contest, war (cf. -क्/ऋत् &c) Naigh. ii , 1
पौंस्ये	MW- पौंस्य is not mentioned in the sense of युद्ध
महाधने	MW- महाधन m. a great contest, great battle ib. Naigh.
वाजे	MW- वाज m. the prize of a race or of battle, booty , gain , reward , any precious or valuable possession , wealth , treasure RV. VS. AV. Pan5cavBr.
अजम्	MW- अजम् is not found in the sense of युद्ध
सद्म	MW- सद्म n. war , battle (= सं-ग्राम (ib.ii , 17
संयत् संयद्	MW- संयत् संयद् f. contest, strife, battle, war (generally found in loc. or comp.) MBh. Ka1v. &c
संवतः	MW- not found in the sense of युद्ध

Table 4. Verification of meaning of words standing for *war*.

3 Special features of Sanskrit wordnet

Verbal concepts: In Hindi wordnet, verbs are not inserted in their root forms. Instead, their dictionary forms like होना *honA* (to be) ,करना *karanA* (to do) ,खाना *khAnA* (to eat) ,पीना *pInA* (to drink) etc. are included in the synset. The last ना *nA* is dropped through suffix stripping in verb morphology and the verb forms are generated using only the initial parts like हो *ho* ,कर *kara* , खा *khA* ,पी *pI* . Sanskrit lexicographers have not conformed to this practice and have inserted the root forms of verbs like भू *bhU* (to be) ,कृ *kR* (to do), खाद् *khAd* (to eat), पा *pA* (to drink), in verbal synsets.

Gender: Sanskrit has grammatical gender. The following practice is followed for tackling the issue of gender in Sanskrit wordnet: (1) In case of nouns all gender variations are included in the synset. (2) Adjectives in Sanskrit have no gender of their own. They take the gender of the nouns which they qualify. Hence in the synset of adjectives only root forms are included. (3) Adverbs- Technically adverbs in Sanskrit do not get conjugated as nouns and adjectives. But, we find that some adverbs have विभक्ति (case ending) suffixes attached to them indicating the closed form of the word in that particular विभक्ति. (case ending). In such cases, they are included as they are, i.e., in the closed विभक्ति form. For example-

(6) सन्निधौ *sannidhau* 'near' (which is actually a locative form of सन्निधि sannidhi), निकटे *nikaTe* 'near' (which is actually a locative form of निकट *nikaTa*), and अदूरे *adUre* 'near' (which is actually a locative form of *अदूर* adUra).

4 Conclusions and future work

One of main challenges in creating the Sanskrit wordnet is dealing with the sheer volume of lexical knowledge accumulated over at least 2000 years. The synsets tend to become long to accommodate coverage of words for a concept. The other challenge is the extremely rich morphology of Sanskrit which produces new words from simple elements. The question of trade-off between a complex morphological interface to the lexical data and the amount of lexicalization needs to be investigated.

The future work is proposed to be carried out in the following directions:

Use of ontology of नव्य-न्याय (Navya-NyAya)

The traditional Sanskrit Texts on Philosophy as well as Medicine contain various discussions on ontological categories and hierarchies. These texts are closely related to the grammar of the Sanskrit Language. The comparison of these ontological structures and hierarchies to the existing one coming from the Hindi wordnet may shed light on new Indowordnet specific issues.

धातु (dhAtu) based WN

There are theories in Sanksrit texts which adhere to the view that all nouns are derived from verbal roots. It is the actions denoted by the verbal roots that can be considered as the base of various objects denoted by nouns. There is a need to test this theory and build a lexical structure where all the verbal roots will be at the nodal level with connected nouns at the leaf level. A brief introduction of this is available in (Kulkarni and Bhattacharyya, 2009).

References

Abhishek G. Nanda. 2009. Tools and interfaces for wordnet construction, linking and maintenance. B. tech project report, Indian Institute of Technology Bombay, Mumbai.

Dan Tufis, Radu Ion, Luigi Bozianu, Alexandru Ceusu, and Dan Stefaescu. 2008. Romanian wordnet: Cirrent state, new applications and proposals. In Attila Tanáces, Dóra Csendes, Vernoika Vincze, Christaine Fellbaum, and Piek Vossen, editors, *Proceedings of the Foruth Global WordNet Conference*:441–445.

H. H. Wilson, editor. 1819. *A Dictionary in Sanskrit and English*. Calcutta.

Krsnaji Govinda Oka, editor. 1913. *Amarakosha of Amarasinha*. Law Printing Press.

Malhar Kulkarni and Pushpak Bhattacharyya. 2009. Verbal roots in the Sanskrit wordnet. In G. Huet, Amba Kulkarni, and Peter Scharf, editors, *Sanskrit Computational Linguistics*, Lecture Notes in Computer Science:328–338, Berlin/Heidelberg. Springer-Verlag.

Malhar Kulkarni. 2008. Lexicographic traditions in India and Sanskrit. *Journal of Language Technology,* (1):160–165.

P. Vossen. 2002. Euro WordNet: General Document. University of Amsterdam.

Raja Radhakanta Deb, editor. 1988. *Shabdakalpa-druma*, volume 1-5. Nag Publishers, 2003 edition. Delhi.

Saranamasimha Sarma. 1968. *Hindi ki Tadbhava Shabdavali*. College Book Depo.

Taranatha Tarkavacaspati Bhattacharya, editor. 2003. *Vacaspatyam*, volume 1-6 of *Chaukhamba Sanskrit Book Series*. Chaukhamba, Banares.

Vishwanath Dinkar Naravane. 1961. *Bharatiya Vyavahara Kosha: Solah Bhasao ka kosha*. Triveni Samgama. [In HIndi.].

Appendix A: Early works on Sanskrit lexical knowledge bases (besides *Amarakosha)*

1.*Naamamaalikaa* of Bhoja (11 C)
2.*SiddhashabdaarNava* of Sahajakirti- (17th C)
3.*Shaaradiiyaakhyaanaamamaalaa* of Harsakirti- (17th C)
4.*Paryaayashabdaratna* of Dhananjaya-Bhatta.
5.*Koshakalpataru*
6.*Naanaartharatnamaalaa* of Irugapa Dandadhinatha (14th C)
7.*Naanaarthamañjarii* of Raghava
8.*DharaNiikosha* of Dharanidas a (12th C)
9.*Shivakosa* of Sivadatta-Misra
10.*Ekaarthanaamamaalaa-vyaksharanamamaalaa* of Saubhari
11.*Paramaanandiiyanaamamaalaa* of Makrandadasa

Appendix 2: Lexicographer's Interface for Sanskrit wordnet building

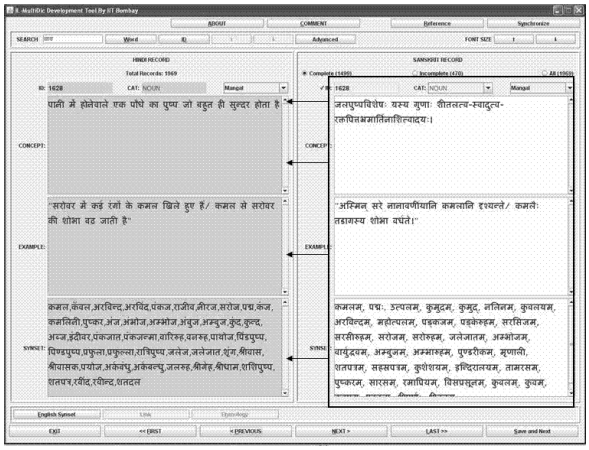

Figure A. 1 Lexicographer's interface.

To create a lexical resource like wordnet, one needs a user friendly tool. Sanskrit wordnet team uses the MultiDict tool developed at the Center for Indian Language Technology, Computer Science Department, IIT Bombay (Figure A. 1). The tool provides an interface for linking the synsets that express the same meaning in different language (Nanda, 2009).

The linker tool (Figure A. 2) is integrated in the interface for cross-linkage between the literals of source and target synsets. It allows a lexicographer to link a literal of the source language to one or more literals in the corresponding target language synset.

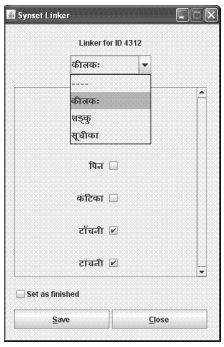

Figure A. 2 Linker

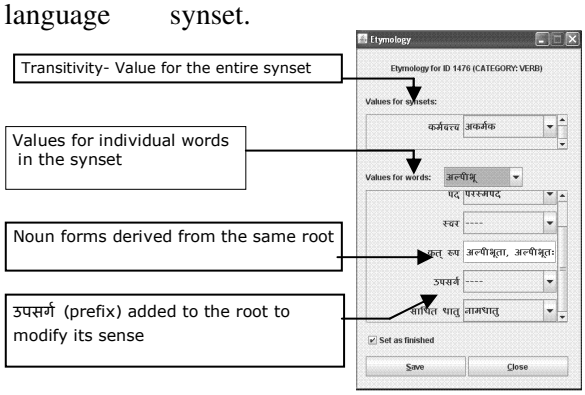

Figure A.3 Morphological elements in the SWN