Introducing Sanskrit Wordnet

Prof. Malhar Kulkarni

Department of Humanities and Social Sciences, IIT Bombay,
Powai, Mumbai-400076, India.

malhar@iitb.ac.in, malharku@gmail.com

Motivation

- Sanskrit …
  - as one of the oldest Indo-European language can provide basis for the linguistic analysis of modern Indo-European languages.
  - as a heritage language needs to preserve the rich lexicon of this language.
  - To preserve the traditional culture and knowledge translation of Sanskrit texts is needed
  - Sanskrit wordnet can be a great help in developing Machine aided translation (MAT) system for Sanksrit.
  - Text Search facility can be developed using the wordlist of Sanskrit wordnet.
  - Sanskrit has a rich lexical tradition but there is no single reference containing all the these lexicas.

Sanskrit language and its relation with modern Indo-Aryan languages

- 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.
- Traditional four-fold division of the lexical units of Indian languages –
  1. तत्सम *tatsama* - words having their origin in Sanskrit and accepted in the modern Indo-Aryan languages without any change in their phonology.
  2. तद्भव *tadbhava*- 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 *deshī*- words which are the native words of the particular language and
  4. विदेशी *videshī* - 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.

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	तुलसी	
	वैष्णवी	तुलसी	
	पावनी	तुलसी	
	पत्रपुष्पा	तुलसी	
{भौंह,भौं,**भ्रू,भृकुटी,** तेवर,कोदंड,कोडंड,अबरू}	**भ्रू**	भौंह	eyebrow, brow, supercilium
	भृकुटी	भौंह	
{**पेशी**, माँस-पेशी, **मांस-पेशी**, माँसपेशी, **मांसपेशी**, माँस पेशी, **मांस पेशी**, नस}	**पेशी**	पेशी	muscle, musculus
	मांसपेशी	पेशी	
{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	

Rich lexical tradition of Sanskrit

Mono-lingual Sanskrit-Sanskrit traditional lexicas that arranged lexical material from the point of view of *synonymy* as well as *homonymy*, –

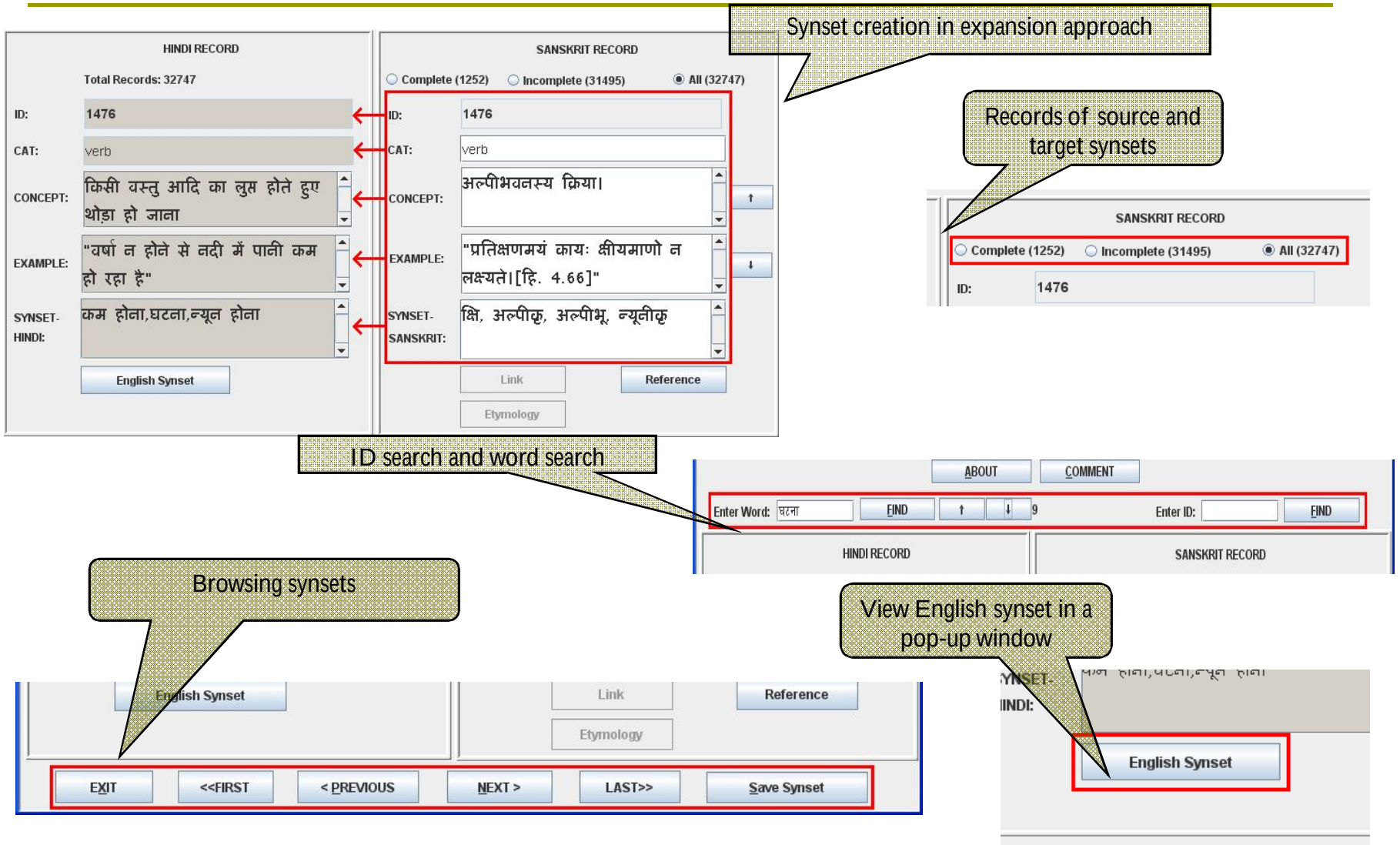
1. *Naamamaalikaa* of Bhoja (11 C)
2. *SiddhashabdaarNava* of Sahajakirti- (17th C)
3. *Shaaradiiyaakhyaanaamamaalaa* of Harsakirti- (17th C)
4. *Paryaayashabdaratna* of Dhananjaya-Bhatta.
5. *Koshakalpataru*
6. *Naanaartharatnamaalaa* of Irugapa Dandadhinatha (14th C)
7. *Naanaarthamañjarii* of Raghava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

- Modern monolingual and bilingual dictionaries-
  - The first modern-day dictionary of Sanskrit was the Sanskrit-English Dictionary compiled by Professor H.H. Wilson and published in 1819 (Wilson, 1819).
  - Monier Williams Sanskrit-English Dictionary (Available online)
  - Two Indian mono-lingual (Sanskrit to Sanskrit) dictionaries-
    - *Shabdakalpadruma* (Deb, 1988) of Pt. Sir Raja Radhakanta Dev and
    - *Vacasptyam* (Bhattacharya, 2003) compiled by Pt Taranatha Tarkavacaspati.
- Sanskrit wordnet mainly relies on –
  - Monier Williams Dictionary
  - *Shabdakalpadruma*
  - *Vacasptyam*
  - Apte's Sanskrit Dictionary

Building Sanskrit Wordnet- Expansion approach

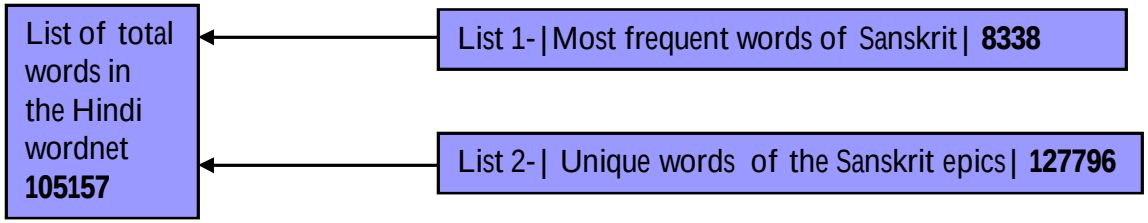
- There are two methods to develop a Wordnet:
  - (1) Expand method- a wordnet is constructed based on an existing wordnet.
  - (2) Merge method- sub-Wordnets for specific domains are built and later merged.
- For Sanskrit Wordnet, the Hindi wordnet is considered as the source resource.
- Though *expanded* from Hindi wordnet, care was taken to ensure that Sanskrit wordnet captures the real lexical structure of Sanskrit language.
- Sanskrit wordnet follows the hierarchy preservation principle (HPP) (Tufis *et al.,* 2008).
  - In the hierarchy of the Hindi wordnet, if synset H2 is a hyponym of synset H1, and the translation equivalents in the Sanskrit wordnet for H1 and H2 are S1 and S2 respectively, then in the hierarchy of Sanskrit wordnet S2 should be a hyponym of synset S1.

Lexicographers interface



Synset creation in the first phase

•Initial word list based approach-



Problems faced:

•Domain based approach (which later extended to all Indian language wordnets): Based on Naravane's भारतीय व्यवहार कोश most common words of Indian languages from following domains are selected.

• Part of Speech distribution: 1512 Nouns, 225 Verbs, 180 Adjectives and 52 adverbs

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	52) Miscellaneous			

Synset creation in the second phase

- Common concepts used in Indian languages are selected from the total Hindi wordnet synsets.
  - Selection procedure-
    - Stage I- All Hindi wordnet synsets were ranked by 6 language resource persons of Hindi, Marathi, Sanskrit and Kannada.

      >>> 13205 concepts are selected.
    - Stage 2
      - Group voting
      - Online voting

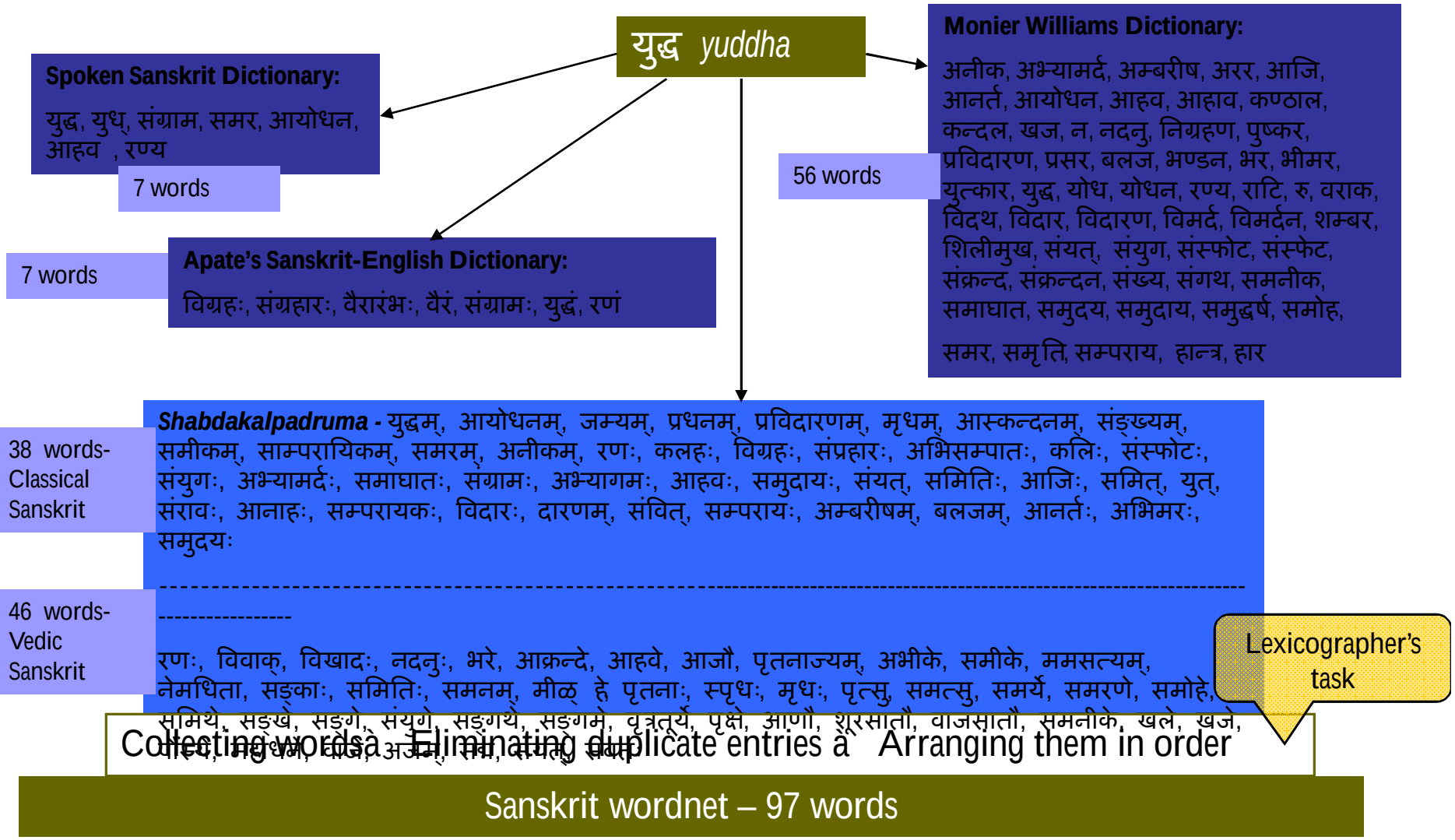
        >>> 10312 concepts are selected.

Synset creation in Expansion approach

- Creating synonymous sets –
- Understanding Hindi wordnet synset.
- Looking for the most common word for the concept expressed by Hindi wordnet synset
  - Collecting synonymous words from traditional lexical resources, modern monolingual (शब्दकल्पद्रुम & वाचस्पत्यम्)and bilingual dictionaries (Monier Williams' Dictionary, Apte's Dictionar and online spoken Sanskrit Dictionary).
  - Arranging words according to the frequency-
    - From most common modern or classical word to the least used Vedic word.


Page 1
Page 2
Head word युद्ध
Gloss
Gloss or etymology?
Set of synonymous words in classical Sanskrit
Problems faced by lexicographers:
1. Large number of synonymous words.
2. Different words used in different periods.
3. Glosses are etymological and sometimes do help in wordnet creation.
4. Examples are either too lengthy (10-15 lines) or too small (not even a complete sentence.)
Objects related to or used in युद्ध
Set of synonymous words in Vedic Sanskrit

Sanskrit Wordnet as a one single reference for researchers and translators.



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*[1] 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*"
- the gloss of a verb in Sanskrit is generally created using technical terms like व्यापार *vyApAra* 'action', जन्य *janya* 'produced,' अनुकूल *anukUla* 'helpful,' etc

[1] Phonological conjoining

Problems faced in Expansion approach

- Difficulty of finding equivalent words:

चाय (Tea)	चाय के पौधे की पत्तियों को पानी में डालकर चीनी, दूध आदि मिलाकर बनाया हुआ पेय पद *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.

चायः (Tea)	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)

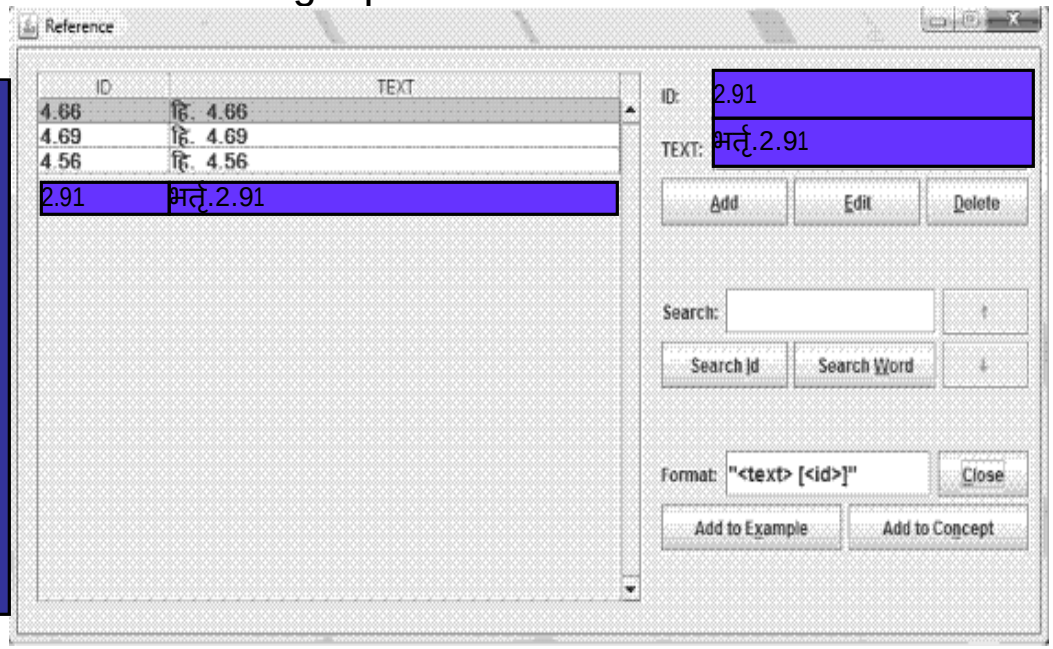
Difficulties with *examples*:

- 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:

The example with the citation is inserted in this format:

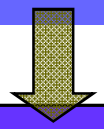
(5) "शशि-दिवाकरयोर्ग्रहपीडनम् ।" [भर्तृ.2.91]
shashi-divAkarayor grahapIDanaM [bhartR 2.91]
(the eclipse of Sun and Moon).

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



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.**



Ø Synset creation

Ø collecting words of all the period

Ø arrange words as per the frequency

[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 problem of meaning attestation:

- Rich lexical tradition.
  - Downside of this fact is that the lexicographer has to verify the consistency of word definitions at every step from multiple sources.
  - For example-To create the synset of युद्ध *yuddha* (War), a lexicographer needs to prepare a record after searching each and every word in various lexica.

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 (= संग्राम) ib.ii , 17
संयत् संयद्	MW- संयत् संयद् f. contest, strife, battle, war (generally found in loc. or comp.) MBh. Ka1v. &c

Special features of Sanskrit wordnet

- Verbal concepts:
  - In Hindi wordnet, verbs are not inserted in their root forms.
  - Instead, their dictionary forms are included in the synset. For example-
    - होना *honA* (to be),
    - करना *karanA* (to do),
    - खाना *khAnA* (to eat),
    - पीना *pInA* (to drink) etc.
  - 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

Challenges

- 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.
- 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

Future work

- **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.
  - 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

- Taranatha Tarkavacaspati Bhattacharya, editor. 2003. *Vacaspatyam*, volume 1-6 of *Chaukhamba Sanskrit Book Series*. Chaukhamba, Banares.
- Raja Radhakanta Deb, editor. 1988. *Shabdakalpadruma*, volume 1-5. Nag Publishers, 2003 edition. Delhi.
- 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.
- Abhishek G. Nanda. 2009. Tools and interfaces for wordnet construction, linking and maintenance. B. tech project report, Indian Institute of Technology Bombay, Mumbai.
- Vishwanath Dinkar Naravane. 1961. *Bharatiya Vyavahara Kosha: Solah Bhasao ka kosha*. Triveni Samgama. [In HIndi.].