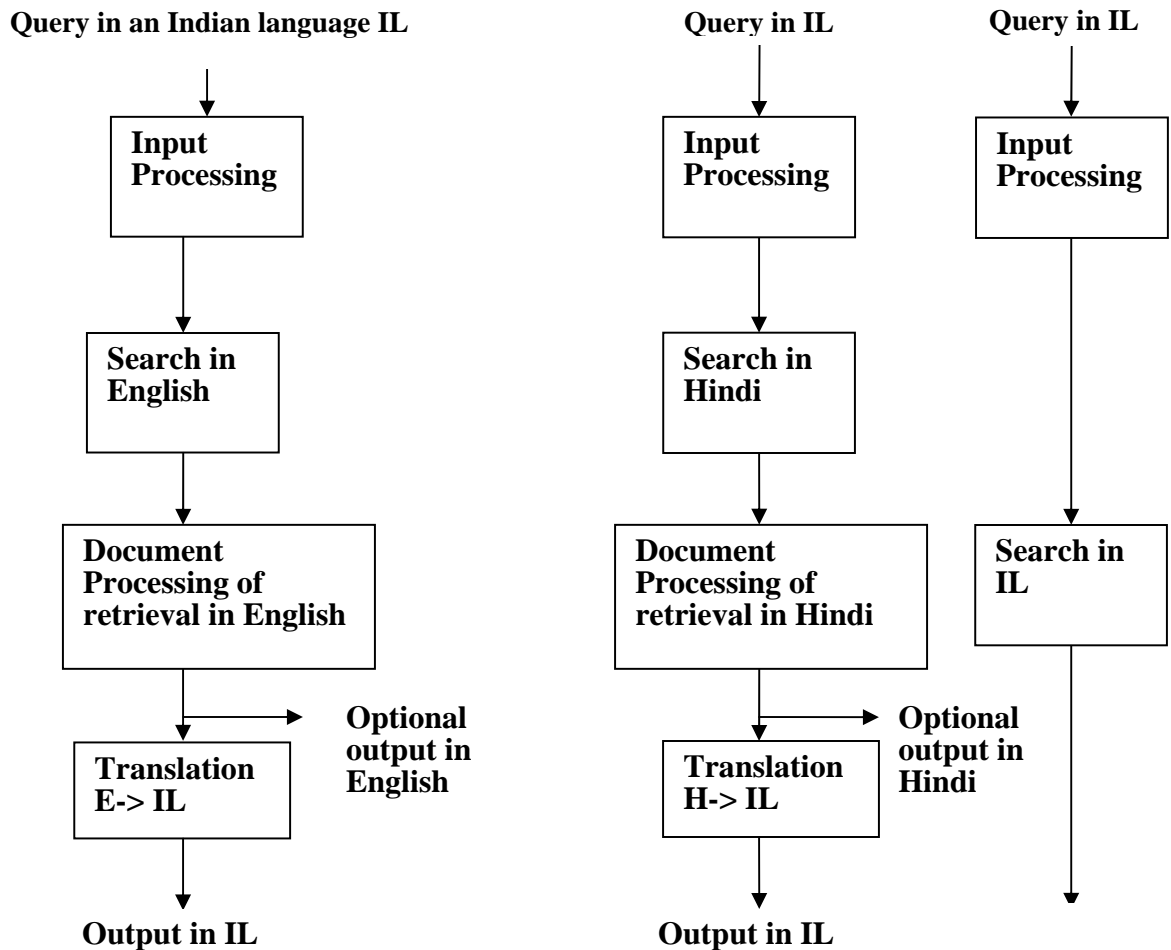


**Executive Summary**

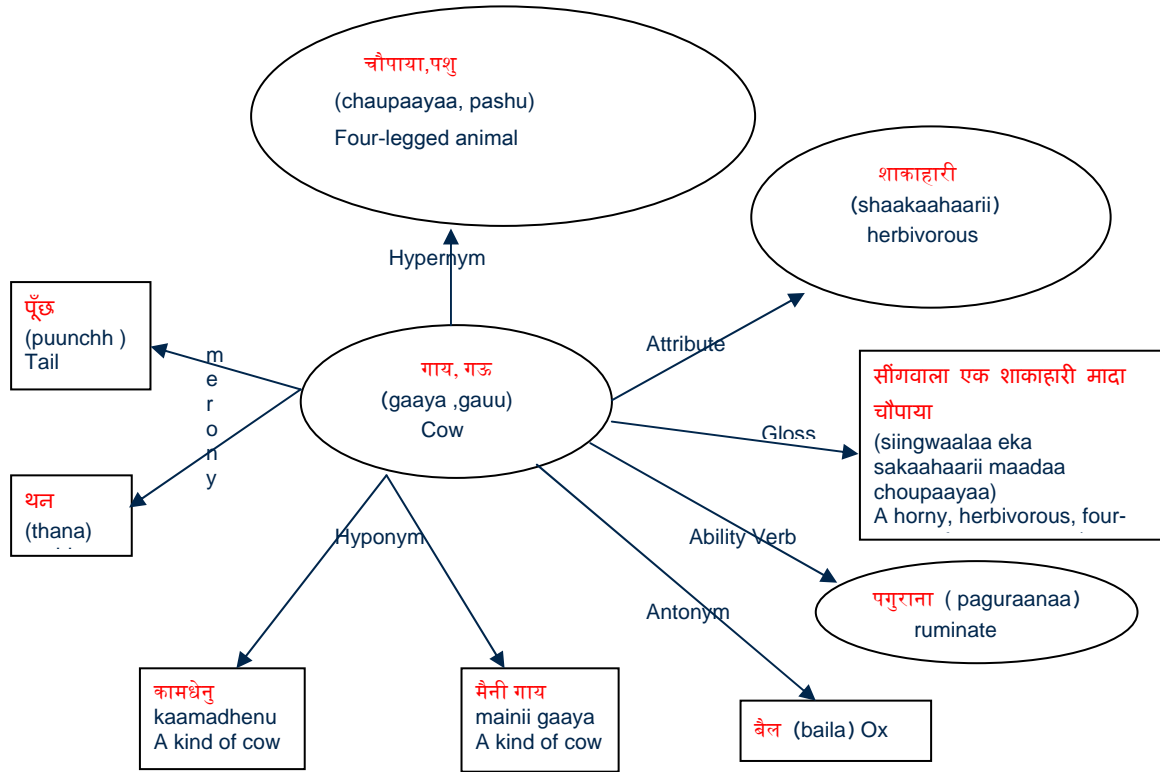
In a multilingual country like India translation between Indian languages as well as between English and Indian languages is a critical task. Similarly critical is the task of Cross Lingual Search where the query is made in an Indian language and retrieval of documents happens in English or Hindi (*vide* Figure 1). All these activities depend on lexical knowledge of high quality and coverage. This lexical knowledge is in the form of *machine readable dictionaries*, *ontologies* (hierarchical organization of concepts) and *wordnets* (a massive graph like structure of words).



**Figure 1: Cross Lingual Search**

Wordnets are lexical structures composed of synsets and semantic relations. **Synsets are sets of synonyms. They are linked by semantic relations like *hypernymy* (*is-a*), *meronymy* (*part-of*) etc.** Wordnets have emerged as crucial resources for Natural Language Processing (NLP). The first wordnet in the world was built for English at

Princeton University<sup>1</sup>. Then followed wordnets for European Languages: Eurowordnet<sup>2</sup>. Since 2000, wordnets for a number of Indian languages are getting built, led by the Hindi wordnet<sup>3</sup> effort at Indian Institute of Technology Bombay<sup>4</sup> (IITB).



**Figure 2: A snapshot of a part of the wordnet**

For an exchange of experience amongst the researchers making wordnet in different languages, and also to chart out a course of action for Pan-Indian wordnet activity, a National workshop on Wordnet creation was organized by the Indian Institute of Technology, Bombay and Amrita University, Coimbatore at Coimbatore from 11<sup>th</sup> June to 14<sup>th</sup> June 2009.

Language groups working in different universities and institutes participated in this workshop. Each language group was a team of trained lexicographers, language experts and computer engineers. Out of 22 official languages of India, the following 13 were involved in this workshop: (1) Hindi<sup>5</sup>, (2) Marathi<sup>6</sup>, (3) Konkani<sup>7</sup>, (4) Sanskrit<sup>8</sup>, (5)

<sup>1</sup> <http://www.wordnet.princeton.edu>

<sup>2</sup> <http://www.ilc.uva.nl/EuroWordNet/>

<sup>3</sup> <http://www.cfilt.iitb.ac.in/wordnet/webhwn>

<sup>4</sup> <http://www.iitb.ac.in>

<sup>5</sup> Hindi/Khadi boli belongs to the Indo-Aryan language sub-group of Indo-European language family. It is a dialect continuum of the Indic language family in the northern plains of India. 2001 census of India noted 422,048,642 speakers of this language. It is spoken in the Indian states and union territories of Bihar, Chhattisgarh, Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh and Uttarakhand.

Nepali<sup>9</sup>, (6) Kashmiri<sup>10</sup>, (7), Assamese<sup>11</sup>, (8) Tamil<sup>12</sup>, (9) Malyalam<sup>13</sup>, (10) Telugu<sup>14</sup>, (11) Kannad<sup>15</sup>, (12) Manipuri<sup>16</sup> and (13) Bodo.<sup>17</sup>

The workshop which was for 4 days discussed the history, speaker population, and properties of the 13 languages involved, and the experience of building wordnets for these languages. This was a highly enriching experience on lexical resources at the national level. The exchange of ideas took place based on **actual hands-on sessions**, where researchers/lexicographers created synsets (the building blocks of wordnets) in correspondence with Hindi synsets. This brought to light the varied ways languages express concepts.

The main conclusion from the workshop was that the development of linked wordnets of different languages of India is a critical activity as far as multilingual information processing is concerned. This end is to be achieved by addressing challenging computational and linguistic problems, like *language divergence*, *large database storage*, *network bandwidth (for online availability of multilingual wordnets)*, *lexical semantics (how to store and process word knowledge)*, *embeddability of wordnet in large applications like machine translation and cross lingual search*.

---

<sup>6</sup> Marathi is an Indo-Aryan language spoken by the Marathi people of south western India and is the official language of the state of Maharashtra. 2001 census of India noted 71,936,894 speakers of this language.

<sup>7</sup> Konkani is an Indo-Aryan language belonging to the Indo-European family of languages spoken in the Konkan coast of India. It has approximately 7.6 million speakers of its two individual languages, Konkani and Goan Konkani.

<sup>8</sup> Sanskrit is a historical Indo-Aryan language and as per the 2001 census of India, there are 6,106 speakers of this language.

<sup>9</sup> Nepali is a language of the Indo-Aryan branch of the Indo-European language family. 2001 census of India records 13,168,484 speakers of this language.

<sup>10</sup> The origin of Kashmiri language is uncertain. According to one view it belongs to the Dardic languages which form a sub-group of the Indo-Aryan languages whereas others believe that it belongs to the Iranian languages. It is spoken in eastern Afghanistan, northern Pakistan, and in the Indian region of Jammu and Kashmir. 2001 census of India recorded 5,527,698 speakers of this language.

<sup>11</sup> Assamese is the easternmost Indo-Aryan language. According to the 2001 census of India there are 13,168,484 speakers of this particular language.

<sup>12</sup> Tamil is the only surviving Classical language in the world and is a Dravidian language. According to the 2001 census of India there are 60,793,814 speakers of this particular language.

<sup>13</sup> Malayalam is one of the four major Dravidian languages of South India. According to the 2001 census of India there are 33,066,392 speakers of this particular language.

<sup>14</sup> Telugu is a Dravidian language mostly spoken in the Indian state of Andhra Pradesh. According to the 2001 census of India there are 74,002,856 speakers of this particular language.

<sup>15</sup> Kannada is one of the major Dravidian languages of India, spoken predominantly in the state of Karnataka. 2001 census of India recorded 37,924,011 speakers of this language.

<sup>16</sup> Manipuri belongs to the Sino-Tibetan family of languages. It is the official language of south-eastern Himalayan state of Manipur, in north-eastern India. According to 2001 census, 1,466,705 speakers of this language are found in India.

<sup>17</sup> Bodo is a Tibeto-Burman language. 2001 census of India records 1,350,478 speakers of this language.

**Acknowledgment:** The funding for the wordnet building activity is primarily from the Government of India. Department of Information Technology (DIT), Ministry of Communication and Information Technology (MCIT) is providing funds for Hindi and Marathi wordnets. Funds for Sanskrit wordnets have been received from Ministry of Human Resources development (MHRD) through the Central Institute of Indian Languages (CIIL). The consortium for wordnet development for languages of the North East is supported by DIT, MCIT. Dravidian wordnet activity and its linkage with the wordnets of Indo-Aryan languages is sponsored by MHRD.

**Introduction: wordnet activity across India**

Wordnet making projects are going on at various institutes for 13 Indian languages mentioned in the executive summary. A number of Indo-Aryan language wordnets are being developed at the Centre for Indian Language Technology (CFILT<sup>18</sup>) at Department of Computer Science and Engineering at IIT Bombay under the leadership of Prof. Pushpak Bhattacharyya<sup>19</sup>.

<b>CFILT, IIT BOMBAY - HINDI, MARATHI AND SANSKRIT WORDNET</b>			
1	Dr.Pushpak Bhattacharyya	Professor	CSE Dept
<b>HINDI WORDNET</b>			
2	Prabhakar Pandey	Research Associate	CFILT (CSE Dept)
3	Mrs.Laxmi Kashyap	Research Associate	CSE / CFILT
<b>MARATHI WORDNET</b>			
4	Gauree Welankar	Research Associate	CFILT (CSE Dept)
5	Ms.Archana A. Naidu	Research Associate	CFILT (CSE Dept)
<b>SANSKRIT WORDNET</b>			
6	Dr. Malhar Kulkarni	Associate Professor	HSS Dept
7	Dr. Irawati Kulkarni	Research Associate	CFILT (CSE Dept)
8	Dr. Chaitali Dangarikar	Ph.D Research Scholar	CFILT / HSS Dept
9	Abhishek Nanda	B.Tech Student	CSE Dept

A separate activity started for a number of languages of the North-East. Following institutes are working with IITB on this:

<b>ASSAM UNIVERSITY- NEPALI WORDNET</b>			
1	Dr. Bipul Shyam Purkayastha	Reader (PI)	Dept. Of Computer Science
2	Alok Chakrabarthy	Ph.D Research Scholar	Dept of Computer Science
3	Mr. Arindam Roy	Asst. Professor	Dept of Computer Science
4	Tek Narayan Upadhaga	Lecturer	Dept of Nepali
5	Dr.Khagen Sharma	Resource Person (Academic)	Dept of Nepali

<sup>18</sup> <http://www.cfilt.iitb.ac.in>

<sup>19</sup> <http://www.cse.iitb.ac.in/~pb>

<b>GAUHATI UNIVERSITY – ASSAMESE AND BODO WORDNET</b>			
1	Dr.Shikhar Kumar Sarma	Reader (PI)	Computer Science
2	Morumi Gogoi	Lecturer & Research Scholar	Computer Science
3	Mane Bala Ramchiary	Lexicographer (Bodo)	Computer Science
4	Sri Biswajit Brahma	Lexicographer (Bodo)	Computer Science
5	Sri Utpal Saikia	Lexicographer (Assamese)	Computer Science
6	Sri Rakesh Medhi	Lexicographer (Assamese)	Computer Science
<b>MANIPUR UNIVERSITY – MANIPURI WORDNET</b>			
1	Dr. Ch.Yashawanta Singh	Professor (PI)	Dept of Linguistics
2	Dr.Hanjabam Surmangol Sharma	Guest Faculty	Dept of Linguistics
3	Yumnam Bablu Singh	Computer Programmer	Dept of Linguistics
4	Bachaspatimayum Preambati Devi	Language Expert	Linguistics

Konkani and Kashmiri wordnets are being developed at the University of Goa and Kashmir University respectively:

<b>GOA UNIVERSITY -KONKANI WORDNET</b>			
1	Dr.Jyoti D.Pawar	Lecturer (Selection Grade) (PI)	Dept of Computer Sc. & Tech
2	Shilpa Desai	Lecturer	Information Technology
3	Ramdas Karmali	Sr. Lecturer	Computer Science & Technology
4	S.S.W.Walawalikar	Consultant	Computer Science & Technology
5	Mr.Sushant K.Naik	Linguist	Konkani Dept & Computer Sc.
6	Damodar Keshav Kaamat Ghanekar	Linguist	Computer Science
<b>KASHMIR UNIVERSITY - KASHMIRI WORDNET</b>			
1	Dr.Aadil Amin Kak	Sr. Asstt. Professor	Linguistics
2	Oveesa Farooq	Senior Linguist	Linguistics
3	Aadil Ahmad Lawaye	Computer Scientist	Linguistics
4	Nazima Mehdi	Senior Linguist	Linguistics

The activity of developing wordnets for Dravidian languages is going on at various universities of south India:

<b>TAMIL UNIVERSITY – TAMIL AND MALYALAM WORDNET</b>			
1	Dr.S.Rajendran <sup>20</sup>	Professor & Head (PI)	Dept of Linguistics
2	R.Amudha	Research Associate	Dept of Linguistics
3	Ramasundari. M	Research Scholar	Dept of Linguistics
4	N.Subha	Research Scholar	Dept of Linguistics
<b>DRAVIDIAN UNIVERSITY - TELUGU WORDNET</b>			
1	Dr.Arul Mozhi	Assistant Professor (PI)	Computational Linguistics
2	Mohammad Ali	Research Associate	Computational Linguistics
3	D.Rajarao	Language Translator	Linguistics
4	M.Kasim Babu	Research Fellow	Computational Linguistics
5	K.Eswar Babu	Research Fellow	Computational Linguistics
<b>AMRITA UNIVERSITY – TAMIL, KANNAD AND MALYALAM WORDNET</b>			
1	Dr. Soman K.P	Professor & Head (PI)	Computational Engg. & Networking
2	Dr.A.G.Menon	Visiting Professor	Computational Engg. & Networking
3	Loganathan	Research Associate	Computational Engg. & Networking
4	Saravanan	Research Associate	Computational Engg. & Networking
5	Dhanalakshmi	Research Associate	Computational Engg. & Networking
6	G.Shiva Pratap	Research Associate	Computational Engg. & Networking

A day wise description of the proceedings now follows:

<b>Day 1</b>	<b>June 11, 2009</b>
--------------	----------------------

The workshop started with a tuneful Saraswati Vandana. Then Prof. Soman Kottipadannayil of Amrita University welcomed the participants to the workshop.

<b>Keynote address: Prof. S.V. Ramanan</b>
--

Prof. Ramanan stated that machines can help understand language. Following developments in the area of computational linguistics are contributing to Natural Language Understanding:

1. Linguistically annotated data: Penn tree bank, POS tagged corpus, NE tagged corpus
2. Wordnets
3. Domain Ontologies
4. Verbnet and Framenet
5. Parsers (Deterministic and statistical)

---

<sup>20</sup> Work on ontology has been going on for Dravidian languages under Prof. S. Rajendran's guidance since a long time.

## 6. Semantic Web

Government of India (GoI) wants all educational information to be available in regional languages. Prof. Ramanan stressed the need for generating properly linked lexical and semantic databases for Indian languages. A wordnet can be one such important tool.

### Multilingual Wordnets: Prof. Pushpak Bhattacharyya

Prof. Bhattacharyya stated that the main aim of this workshop was to construct the **Indowordnet**.

In 2002, the Department of Information Technology initiated the project called Technology Development for Indian Languages<sup>21</sup> (TDIL) with the objective of developing basic language and information processing tools and techniques. One of the many benefits accruing from this endeavour was Hindi and Marathi wordnets developed at IIT Bombay.

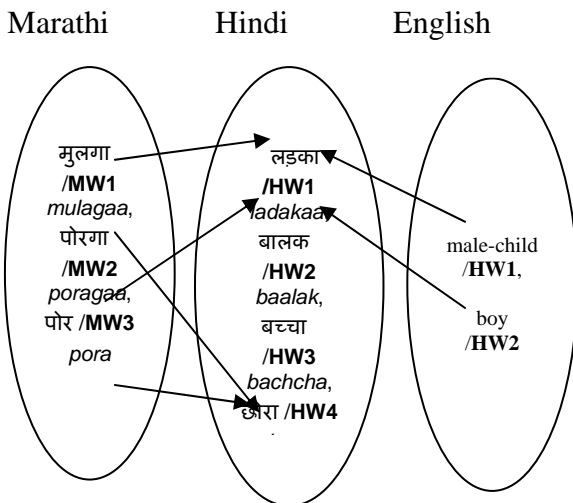
#### Wordnet Development at IIT Bombay:

Hindi Wordnet was created from the **First Principle**. This means words of Hindi were taken one by one and were examined for their senses. For each sense a synset was created. After that the synsets were linked with semantic relations.

Wordnets creation for Marathi, Sanskrit, North-Eastern languages, Konkani and Kashmiri have been started using the **Expansion Approach**. In the Expansion Approach, Hindi Wordnet synsets are understood by the lexicographer and the corresponding target language synsets expressing the same sense are created.

Marathi wordnet<sup>22</sup> was the first one to be developed using this expansion approach and semantic relations were borrowed from Hindi wordnet which saved a lot of time.

#### Dictionary Standardisation:



**Figure 3: Cross linked synset members for the concept: a youthful male person**

<sup>21</sup> <<http://tdil.mit.gov.in/>>

<sup>22</sup> <http://www.cfilt.iitb.ac.in/wordnet/webmwn>

Prof. Bhattacharyya also mentioned that a Multilingual Dictionary framework has been created which links *synsets* of languages, and within the synsets, the words (*vide* Figure-3). This unique framework would facilitate work on machine translation, multilingual word sense disambiguation and Cross Lingual Search.

#### **Hindi Wordnet: Prabhakar Pandey and HWN Team, IIT Bombay**

In his presentation Mr. Prabhakar Pandey explained the process of Hindi wordnet creation. Words are taken and for each word and for each sense, synonyms are found and inserted in the synset. While creating a synset, three principles, (1) *Minimality*, (2) *Coverage*, and (3) *Replaceability* are followed rigorously.

Mr. Pande touched upon the semantic relations between synsets, and also the ontology backing up these synsets and lexical relations between the synset members, *i.e.*, words. The total number of synsets in the Hindi wordnet is 32950, covering 77800 unique words. All the 32950 synsets are semantically linked. Out of these 32950 synsets, 13830 synsets are linked with the synsets of the Princeton wordnet.

#### **How to construct a new WN: Prof. Pushpak Bhattacharyya, IIT Bombay**

In this lecture, Prof. Bhattacharyya explained the difference between creating a new dictionary and creating a new wordnet. In wordnet creation, the focus shifts from words to concepts. Instead of selecting common words, we select the most common concepts. For example, सूर्य (Sun), पृथ्वी (Earth), जल, पानी (Water) *etc.* are very common concepts. After selecting a concept, all the words standing for that concept are stored as the set of synonymous words.

Prof. Bhattacharyya further explicated the principles of wordnet construction. *Minimality* principle insists on capturing that minimal set of the words in the synset which uniquely identifies the concept. For example {*family, house*} uniquely identifies a concept (*e.g.* “*he is from the house of the King of Jaipur*”). *Coverage* principle then stresses on the completion of the synset, *i.e.*, capturing ALL the words that stand for the concept expressed by the synset (*e.g.*, {*family, house, household, ménage*} *completes the synset*). Within the synset the words should be ordered according their frequency in the corpus. *Replaceability* demands that the most common words in the synset, *i.e.*, words towards the beginning of the synset should be able to replace one another in the example sentence.

#### **Selection of most common concepts:**

To select the most common concepts from approximately 32000 synsets of HWN, following steps were taken:

- a. Initially 32000 synsets were distributed among 6 people. Each one ranked them into 4 categories, *viz.*, (i) *Common*, (ii) *Common in Indian languages* (iii) *Common in Hindi* and (iv) *Uncommon*, with the help of a specially designed tool for synset ranking. By this process 16000 synsets were filtered in.
- b. These 16000 synsets were again ranked by voting. This process selected 11000 synsets as common synsets.
- c. An online interface was provided to rank these 11000 synsets by the NLP group at IIT Bombay.

- d. भारतीय व्यवहार कोश (*bhaaratiya vyavahaara kosha*) compiled by D. N. Narwane was used to create a set of core concepts necessary for everyday living and communication. 2000 synsets were selected as core synsets and distributed to other language groups.
- e. Rest of the common synsets were also distributed, but these would be linked only after finishing the 2000 core concepts.

### **The Offline Linking Tool: Tutorial by Abhishek Nanda, IIT Bombay**

Abhishek Nanda, a B.Tech student of IITB presented a java-based off-line tool which is used for synset creation using the expansion approach. This tool was made available to various language groups. The tool now comprises of a word search feature, an English synset viewer, a quote reference interface, an etymology data interface and a cross linking interface.

Following suggestions were made by the users of the tool:

1. Make additions to the Etymology feature: this was mainly in the context of Sanskrit wordnet.
2. Add functionality to set up lexical relations within synsets (Antonymy and Gradation).
3. Access to the ontology nodes and semantic relations like meronymy-holonymy in the off-line tool.
4. Correct font problems for Marathi, Kannada and Assamese (some characters like eyelash र् in a Marathi word तन्हा are not encoded correctly in the font used by the tool).
5. For every multi-word expression in the synset field, replace the spaces between the words with underscores (eg. *reach\_in* for *reach in*)
6. Parallel corpora tags should be added in the example field, so that when the examples are translated from the Hindi wordnet to the target language wordnet, they will have the “parallel” tag attached to them. By doing so, it will be easy to include translated sentence in the parallel corpora of Indian languages.
7. Keyboard shortcuts for find (ctrl-f), undo (ctrl-z) *etc.* should be added. Some participants also demanded a provision of copying the entire set of synsets with the keyboard shortcut.

### **Experience of different language groups for linking common synsets**

The participants of the workshop were given the task of linking the synset of सूर्य (Sun) to synsets of various languages. Each group expressed its opinion regarding the HWN synset and its experience of linking with it. There was a serious discussion regarding the following two fields of a synset:

1. Concept field, *i.e.*, *gloss*: Some language groups found that the concept field in (the) HWN synset (हमारे सौर जगत का वह सबसे बड़ा और ज्वलंत पिंड जिससे सब ग्रहों को गरमी और प्रकाश मिलता है; *the biggest, burning ball of our solar system, from which all planets receive heat and light*) was highly technical, and from the pedagogical point of view it was difficult to understand this concept. Some other groups, however, found the definition very precise, and they created a

- similar definition. It was pointed out, though, that **there is circularity in the definition, in the sense that the term “solar” embeds in it the notion of “sun”**.
2. Example field: There are two possibilities in which any language group can fill the example field of any synset (a) creating its own example and (b) translating the HWN example. Some language groups found that translating the HWN example into their own language was less time consuming, whereas other groups found it producing an unnatural sentence in their language.
  3. Some groups from Dravidian languages and North-eastern languages face difficulty in understanding Hindi. They were asked to use the EWN synset.
  4. It was also noticed that some of the HWN synsets from the most common 2000 synsets are not linked with the correct English synset. IITB will set up correct linkages as early as possible and a fresh set of source files will be distributed along with the modified tool.
  5. The HWN group also suggested to the other groups to follow certain conventions while preparing the synsets. They are:
    - a. No comma [,] or full stop [.] should be added at the end of the set of synonymous words in the synset field.
    - b. Whenever a language group provides more than one example in the example field, these two examples should be separated by a slash [/].
    - c. If a lexicographer is putting any sentence from well known reference books in the concept or example field, the reference field should be used to cite the source of the text.

Regarding the synset or concept of सूर्य (the Sun), it was observed that the Indo-European languages have more words (approximately 20) for the concept than the Dravidian languages (typically, 5-6).

**Day 2**

**12, June 2009.**

### **Marathi Wordnet and Synset linkage: Mrs. Gauree Welankar, IIT Bombay**

In her presentation Mrs. Welankar informed that the Marathi wordnet (MWN) was the first wordnet to be developed with the expansion approach from HWN. There are 21,540 synsets in the MWN consisting of 42,000 unique words. Ms. Welankar suggested that orthographic variations like आंबा (*aambaa; mango*) and अंबा (*ambaa; mango*) be included in the synset. Before saving a synset, its ID, its semantic relations and the part of speech should be checked<sup>23</sup>. Regarding the linkage of words from within a synset to the words of another language's synset, the principle followed is that the most frequent word of the source language should be linked to the most frequent word of the target language.

### **Foundational Discussions**

Regarding the translation of examples, many participants argued that if these examples were not natural in their languages, then users of wordnet would be put off. Some freedom should be given to the lexicographer for constructing examples.

<sup>23</sup> This checklist was related to the old off-line tool. Some points may or may not be applicable to the new off-line tool.

However, the issue of standardization versus the creative freedom of the lexicographer was raised by Prof. Shikhar Sharma of Guahati University and other participants. It was suggested that if a language does not have a word for a concept (typically happens for culture specific situations), a lexicographer should adopt the following steps:

1. Transliteration
2. Multiword expression (short phrases)
3. Coining of new words

in that order of priority.

Regarding region specific and culture specific words, the general policy adopted by HWN and the MWN was to assign a specific ID range for such concepts. However, this needs synchronization among lexicographers. Consider, for example, नऊवारी (a special kind of nine yard saree which women in Maharashtra, Goa and in other states wear). This synset for this concept is already created by MWN and a specific range of IDs has been assigned to it. However, the Goa group, which started very recently, would not know the existence of this synset and would duplicate this.

ID Range	HWN	MWN	Konkani	Assamese	Tamil	Telugu
00000	<b>Common synsets across the languages</b>					
50000						
		Marathi Specific Concepts i.e., नऊवारी				
60000						
			Konkani Specific Concepts i.e., नऊवारी			
70000						

**Table 1: Id ranges for languages**

Some participants suggested that assigning of IDs to language or region specific concepts (*vide* Table 1) should be managed by the HWN group. However, Dr. Pushpak Bhattacharyya suggested that a multilingual group should be formed which would monitor the addition of these language and region specific concepts and would assign IDs to them.

#### **What is a Wordnet?: Dr. Pushpak Bhattacharyya**

Prof. Bhattacharyya explained that a wordnet is a dictionary based on psycholinguistic principles<sup>24</sup>. Human lexical memory for nouns is organized as a hierarchy.

<sup>24</sup> Fellbaum, C. (ed.) 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

**Lexical and Semantic relations:**

Dr. Bhattacharyya elaborated that lexical relations like antonymy are relations between words of the language, whereas semantic relations like hyponymy-hypernymy are relations between concepts.

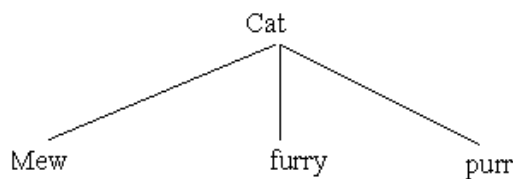
**Lexical relations**  
 $Word_1 \leftrightarrow Word_2$   
**Lexical Matrix: Polysemy and synonymy**

**Semantic relations**  
 $Concept_1 \leftrightarrow Concept_2$

Word Meanings	Word Forms			
	F1	F2	F3	Fn
M1	E1,1	E1,2		
M2		E2,2		
M3			E3,3	
Mm				Em,n

**Figure 4: Explains the way concepts and words relate to each other and naturally account for the existence of synonymy and polysemy.**

When we think of including relations between words, a fundamental design question that arises is *how should the following two types of relations- paradigmatic and syntagmatic- be captured in the lexical database.* Syntagmatic relations are relationships between words which arise from the occurrence of the words in the same syntagma, *i.e.*, same sentence or phrase. For example:



Thus “cat” and ”mew” occur together in a sentence, as do “furry” and “cat, and “purr” and “cat”.

Paradigmatic relations are the relations between words which can in some way substitute one another, and may or may not co-occur, for example,  $Cat \leftrightarrow Mamal$ . Most of the lexical information encoded in a wordnet is of paradigmatic nature.

For English, about half of the associated words are *syntagmatically related* and half are *paradigmatically related*.

A wordnet has a fundamental role in word sense disambiguation and sense tagging. Words in natural language are polysemous. However, when synonymous words are put together, a unique meaning emerges. This principle is known as principle of Relational Semantics.

Componential semantics is the alternative to relational semantics. In *Componential Semantics* each word is considered as a bundle of semantic features. Suppose we were to distinguish between the two senses of the word *CAT*. This word has two senses: ‘*cat*’ as an animal (*Cats like milk*) and ‘*cat*’ as an expert (slang; *this guy is a cat in chemistry*). The table below shows the componential semantic representation of the two senses of ‘*cat*’.

	Animate	Human	Moving	Carnivorous
CAT as animal	Y	N	Y	Y
CAT (slang) in the sense of an expert	Y	Y	Y	Y

**Table 2: Example of Componential Semantics:**

The two tuples for the two meaning of ‘*cat*’ (Table 2 above) differ in the feature ‘human’.

### Semantic Relations between synsets:

Semantic relations link concepts expressed by synsets. For example the concept of *lion* finds its generalization in the concept of *animal*. These relations are foundations of computational semantics. Newer versions of wordnets are proposed to have the *metonymy relation* to capture metaphor, e.g., *Lion* in one sense means *animal* and in a metaphorical sense means *bravery*. Main lexico-semantic relations in wordnets are the following:

1. **Hypernymy-Hyponymy:** This relation is between word senses and denotes *generalization (hypernymy)* and *specialization (hyponymy)* respectively. The relation is transitive and anti-symmetric.  
All language groups will be using the same synset IDs as those of HWN. This will ensure that semantic relations get set up automatically.
2. **Meronymy-Holonymy:** This relation stands for *part-whole* relationship between synsets. This also is borrowed automatically from HWN.
3. **Antonymy:** This relation is selective about word forms and is always set up between words, and never between synsets. So this relation cannot be borrowed from the HWN, but needs to be created.
4. **Troponymy:** This relation refers to the manner of some action. It has temporal inclusion. For example: *Limping is a manner of walking*. This is also between synsets and gets borrowed from the HWN.
5. **Entailment:** This relation implies implication between verbs. Proper inclusion of activity is expected. *Snoring implies sleeping*. This is a synset linking relation and is borrowed from HWN.

6. Gradation: This stands between word forms. It defines a spectrum of antonymy, *e.g. cold-warm-hot*. This lexical relation cannot be borrowed from HWN and needs to be set up.

### **How to capture nuances?**

Having nuances in a language is not unusual. These nuances have fine grained meaning distinctions, *e.g.*, we have different verbs for different kinds of dying: मरना, स्वर्गवासी होना, शहीद होना, परलोक सिधारना *etc.* Here it is necessary to capture the granularity of meaning.

### **How much of morphology to be introduced in WN?**

1. Inflectional Morphology: These forms are not entered in the WN.
2. Derivational Morphology: Many times derivations change word meaning (*e.g.*, *read-reader*). Derived words are entered in the WN.

If the morphology entails a meaning change which is non-productive then it has to be lexicalised. This gives rise to a question of what is productive and what is non-productive? This is an everlasting question in dictionary creation.

### **Dravidian Wordnet: Prof. Rajendran and Amrita University team.**

The concept of a Dravidian WordNet had emerged in a Workshop (2-3 June, 2003) on WordNet for Dravidian Languages organized at the AU-KBC Research Centre, in collaboration with the Central Institute of Indian Languages and Tamil University.

A foundational Ontology has been used for the preparation of the wordnet. Dr. Rajendran observed that the Dravidian wordnet was a natural constituent of the Indo-WordNet. Dr. Rajendran also explained the design of individual wordnets.

The issue of whether to use standard words (साधू usage) or to use local words while creating synsets was discussed.

### **Day 3**

**June 13, 2009**

### **Telugu Wordnet: Dravidian University team**

The team informed that the Telugu language belongs to the Dravidian family of languages. It has a literary history of over 3000 years. Telugu has 4 dialects, *viz.* *Puurva*, *Maddiya*, *Uttara*, and *Dakshina*. It has a highly Sanskritized vocabulary and also has some borrowings from Persian, Arabic and Urdu. Dr. Arulmozi described their wordnet activity and also suggested some modifications in the MultiDict tool developed at IIT Bombay used for synset linking.

### **Tamil and Malyalam Wordnet: Tamil University and Amrita University**

It was informed that the Tamil Wordnet currently has 50K roots. The convenience of constructing the Tamil wordnet from the Malayalam wordnet was mentioned, underlining the importance of *expanding* from a closely related language.

### **Sanskrit Wordnet: IIT Bombay**

Prof. Malhar Kulkarni discussed the importance of the Sanskrit wordnet and also highlighted the specific principle of योग्यता (*yogyataa; fittingness*), on the basis of which one can disambiguate a sentence. He informed that the Sanskrit wordnet is being developed using the expansion approach. It has created more than 1700 synsets. Prof. Kulkarni focused on some additions made to the SWN, like the features of storing etymology, reference and extra information for morphological analyzer.

### **Konkani Wordnet: Goa University**

The presentation dealt with the history of the Konkani language, its morphology, gender and number system, pronunciation and some observations regarding common synsets. Konkani is derived from Sanskrit through *Prakrit*, it is said. There is also a view that it is derived from *Apabhramsh* and *Mundari*. The team discussed the following two problems:

- a. Unicode does not support some of the characters like eyelash र् (ऱ्य), which are also used in Marathi.
- b. One tense known as *unknown past* or *inferential past* is a specific feature of Konkani which may not be captured by other languages.

### **Kannada Wordnet: Amrita University**

Kannada Wordnet team focused on the survey of their work on NLP. They also discussed a few example synsets created in the workshop.

### **Manipuri Wordnet: Manipur University**

The Manipuri Wordnet team discussed the history of the Manipuri language, its script, its phonology, morphology and syntax. They also mentioned the following challenges:

1. No comprehensive good dictionary is available for Manipuri.
2. No dictionary of synonyms for Manipuri has yet been compiled.
3. subtle differences in the manner of action, *e.g.*, different verbs to indicate different types of cooking.
  - a. *phut-pa* ‘to cook vegetables with water only’
  - b. *ngaan-ba* ‘to cook (usually vegetables) in steam’
  - c. *taaw-ba* ‘to deep-fry’

The problem here is to decide whether the verb for ‘cooking’ should form a single synset or should each verb belong to a different synset?

### **Nepali Wordnet: Assam University and Team**

The Nepali Wordnet group focused on the language specific features and its structure. They also discussed their User Interfaces and tools developed for synset linkage.

### **Kashmiri Wordnet: Kashmir University and Team**

The Kashmiri Wordnet group presented specific features of their language, problems and issues related to synset creation and linkage. They also suggested some modifications in the current HWN synsets and stressed the need for support in the linkage tool for *right to left* scan within a synset.

One interesting problem mentioned was that different words exist in different religious communities for the same concept. For example, potable water is called *tresh* and water in general is called *aab* by the Muslim community in Kashmir, whereas Hindus use the same term *pon* for both. This raises the question of what the synsets for water should be in Kashmiri wordnet.

### Some issues of common concern across Languages

#### Homography and Homophony:

It was mentioned that the lexicographer must be aware of the homographic and the homophonic words of a language. The word आम (*aam; mango*) means a fruit and a tree of that same fruit in Hindi. Konkani has a single un-compounded word for आम as a fruit, but does not have a single un-compounded word for आम as a tree. In such cases, *small expressions, compounds or multiword phrases* should be included in the corresponding synsets.

This gave rise to the question of how much of compounds and multiword phrases should be there in a WN? If there is no expression for a particular concept, then the lexicographer should construct a multiword expression to paraphrase that concept. For example, Konkani could introduce some expression like, पेराचा झाड for आम वृक्ष (Mango tree). While introducing such multi-words the lexicographer should separate component words by an underscore ( \_ ).

Some languages have a single word and a multiword expression for the same concept, *e.g.*, in Nepali, मोह and मोह-माय both meaning *infatuation*. To adhere to the principle of coverage, a lexicographer should include both expressions in the synset. However, uncommon expressions can be omitted. Also should be omitted multiwords for which the second component is meaningless, like चाय-शाय (tea and snacks), रोटी-शोटी (*bread*s).

If a target language multiword expression is more than 5 words long, the lexicographer should transliterate the source language word.

Inflected forms should not be stored in synsets.

In Sanskrit many compounds like कनकप्रभ, सुवर्णप्रभ (both meaning “bright like gold”) get created by a *productive process*. In such cases only the most common representative word from the class will be chosen. That is, सुवर्णप्रभ should be stored if it is the most common word representing the concept of *bright like gold*.

The lexical principle behind this is that one should not store productive forms. For example, an English dictionary does not store all the forms formed by the suffix “-er” like

‘reader’, except where there is a change in the meaning. Then, in the same way, is it possible to omit words like वाचक, लेखक, चालक which have “-अक” suffix to mean *agent* of a particular action.

A critical issue discussed was that of **common ontology backing up each language’s wordnet**. As of now, the HWN and the Dravidian wordnets have different ontological structures. The possibility of a uniform ontological structure should be looked into. To make this possible, a request was made to Prof. Rajendran to distribute the ontology developed by him. The IndoWordnet group would make a study of the DWN ontology and see the possibility of backing up each and every synset by the nodes of this ontology.

**Day 4:**

**June 14, 2009**

#### **Assamese and Bodo Wordnets: presented by Guahati University**

The Assamese wordnet group described their strategies for building a wordnet, the challenges in the creation of synsets and their future plans. They also brought up the issues of standardization and suggested that the coining of any new word by a group be broadcast to all the members. They presented a brief history of Assamese and Bodo languages, the scripts used for these languages, their phonological characteristics and their morphology.

It was informed that the Bodo vocabulary contains approximately 15,000 to 20,000 words. This was not sufficient to express all the synsets from Hindi using only Bodo words. At the same time indiscriminate borrowing of words was not an acceptable option. It was suggested that Bodo wordnet would form small phrases for lexicalizing concepts. That failing, they can transliterate Hindi or English words.

An observation was made that the Assamese group sometimes gave two example sentences. They were requested to use forward slash [ / ] to separate examples.

#### **WSD in a Multilingual Setting: Dr. Pushpak Bhattacharyya**

All words WSD in a multilingual setting is a highly challenging task. WSD can be performed using one of the three approaches: *knowledge based*, *machine learning based* and *hybrid*. IITB’s work on WSD uses a hybrid approach with parameter reuse/projection. The parameters are:

1. Domain specific sense distribution of words
2. Dominant concepts in the tourism domain: {*place, country, city, area*, e.g. }
3. Corpus co-occurrence frequency of senses
4. Conceptual distance between nouns
  - a. Wordnet hierarchy plays an important role in this. The length of the path between the noun concepts help in disambiguation
5. Relations learnt from semantic graph obtained from the wordnet hierarchy

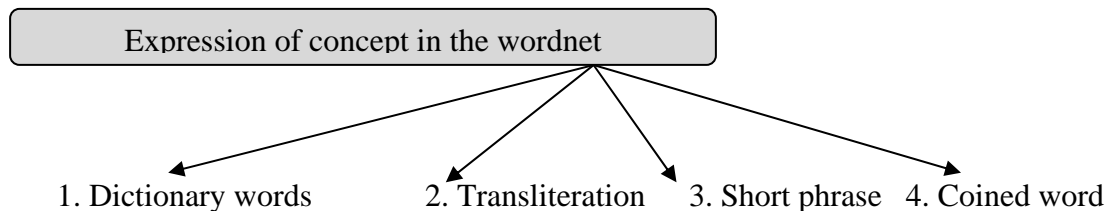
Multilingual dictionary (*vide* Figure 4) facilitates projecting these parameters.

A novel sense scoring function inspired by the energy function of the Hopfield separates well the self merit (expressed as the sense distribution value) of the synsets from their interaction with the senses of other words in the sentence.

## Concluding Session:

Following points were made in the concluding session of the workshop:

1. Wordnet's central concern is to express a concept unambiguously. If there were numbers and/or pictures to express the concepts, that would have been the ideal situation. However, to express concepts with a set of word (s) we can follow the options mentioned below:



2. Dictionary words are included in the wordnet according to the frequency of their use. Options 2, 3 and 4 should be used with discretion, respecting the native speakers' sensitivities.

As for including newly coined words (for example, हिमदुग्ध, *himadugdha* meaning *ice cream*, coined by some author in Marathi), it was felt that *Standardization* may be a problem. Coining of new words should be avoided till the method of coining and the procedure of standardization are decided. Some ways of standardization were proposed but there was no consensus. One view was to validate the words by keeping them on the web and asking for opinions.

At this stage of the project, we have to essentially work on common concepts, and hence there may be no need to coin new words. If a word has already been coined by a government body then it should be accepted. The words should be searched in the terminology dictionaries made by the government.

3. It was stressed that for better results in natural language processing and machine translation the following three things are very important:
  1. Good lexical resource or structured dictionary database
  2. Algorithms to capture linguistic phenomena
  3. Tools/Interface.
4. Prof. Bhattacharyya announced that the **5<sup>th</sup> Global Wordnet International Conference will be held at IIT Bombay on January, 31 to February 4, 2009**. The event will be attended by top researchers and developer from the wordnet and NLP community. He urged the participants to submit papers to the conference and also to participate actively in the event. The website of the conference is at <http://www.cfilt.iitb.ac.in/~gwc2010/index.php>.

### Plan of Action:

1. Errors in the Hindi-English synset linkage would be modified at the earliest. The synsets with modified and correct linkage would be distributed to all the language groups along with the modified wordnet linking tool.
2. The deadline for completing the initial core 2000 synsets is Sept. 2009.

3. After the completion of core synsets, a similar workshop would be organized by Guwahati University.
4. A Google group, called IndoWordnet would be created to discuss the problematic issues and report monthly progress (*this is already done*).
5. IndoWordnet Wiki would be created for documentation (*already done*).
6. The IITB HWN team will take initiative in visiting and coordinating the wordnet work of language groups across India.
7. Like the off-line tool, the MYSQL database which stores words and word properties should be shared.
8. Ontological linkage from Dravidian wordnet to the HWN is to be explored.
9. Uniformity of the database structure and tool has to be maintained.
10. **Same synset ID has to be maintained across languages.**

**Directions for synset creation:**

**On Gloss:**

- Gloss will be based on the HWN or EWN concept fields.

**On Example:**

- HWN sentences should be translated. If they are unnatural in the target language, only then construct a new example.
- Example sentences should be such as to explicate the concepts well.

**On the Set of synonymous words:**

- The three principles of synset creation - minimality, coverage and replaceability must be followed meticulously.
- Care should be exercised not to include hyponymous or hypernymous words in the synset.

The workshop concluded with a vote of thanks for Prof. Soman and his team in the Amrita University for all the logistics and the local support provided. The IITB wordnet team was thanked for its initiative and the sharing of the expertise and experience in wordnet building.