

**Report of the Indo-WordNet Workshop
14-16 January, 2003**

Central Institute of Indian Languages

(Ministry of Human Resource Development, Department of Secondary
and Higher Education, Government of India)
Manasagangotri, Hunsur Road
Mysore-570006

Recognizing the immense importance of lexical resources, the Indian languages WordNet workshop was jointly organized by the Central Institute of Indian Languages (CIIL) Mysore and Indian Institute of Technology (IIT) Bombay from the 14th to the 16th of January, 2003. The objective of the workshop was to explore methodologies for constructing the WordNet for Indian languages and then linking them internally to produce the Indo-WordNet, which eventually would be linked to the English WordNet and the Euro-WordNet (a conglomeration of European languages' WordNets). It is now an accepted fact that no meaningful research and development in language processing, information extraction and machine translation can be carried out without WordNets.

In India, WordNet building activities are going on for Hindi and Marathi at IIT Bombay, Tamil at Anna University-K.B Chandrashekhar Research Centre (AU-KBC) Chennai and Tamil University Tanjavur, Gujarathi at MS University Baroda, Oriya at Utkal University Bhubaneswar and Bengali at IIT Kharagpur. The Hindi WordNet is at an advanced stage of development with about 11000 semantically linked synsets and with the associated software and the user interface.

In the workshop all the major languages of India were represented. The need for utilizing the enormous amount of linguistic work in the country for the purpose of WordNet building was stressed and explored. It was recommended that a website where related information, software and resources will be kept in a browsable and freely downloadable form should be set up.

Tutorials on the fundamentals, methodologies and the applications of the WordNet were held. The concepts of (i) Synsets, (ii) Semantic Relations and (iii) the Interface of the WordNet were explained. Since the synsets are the building blocks of the WordNet, considerable amount of time was spent on describing the structure, principle of creation and the associated parts of a synset.

The gloss plays a very important role in the WordNet since it is through this that the synsets are linked across WordNets. Thus, in the Indo-WordNet, the language specific WordNets are expected to have identical glosses and examples as far as possible. The advantage of this is the possibility of creating multiway parallel corpora.

The synsets should be constructed abiding by the three principles of

- (i) Minimality (the minimal set of words to make the concept unique)
- (ii) Coverage (The maximal set of words- ordered by frequency in the corpus- to include all possible words standing for the sense)
- (iii) Replacability (The example sentence should be such that the most frequent words in the synset can one another in the sentence without altering the sense)

There was demonstration of the application of the WordNet in automatically creating document for specific dictionaries. The words obtain their disambiguators and semantic attributes from the WordNet. The other method of constructing WordNets using an ontology- motivated by Nida's concept classification was discussed for developing the Tamil WordNet. It is created through the following steps: (i) extraction of words from the dictionary, (ii) grouping of words into domains and sub-domains and (iii) arranging the groups hierarchically.

After this, all the participating language groups exchanged notes on an exercise done on a 100 synset sample from the Hindi WordNet provided by IIT Bombay to all the language groups 3 months prior to the workshop. These 100 synsets cover all the parts of speech and were from major conceptual categories like natural object, action, quality etc. It was interesting to observe how words assume different shades across languages, how the glosses become tricky to create for commonly used terms, how words prefer collocations, how example sentences often are directly adaptable with minimal changes from one language to another "close" language. This was an extremely educative experience which clarified the methodology of WordNet construction. While this exercise was going on it was realized that the following need discussions:

- (a) The ontology behind the WordNet.
- (b) Compositional approach to the construction of the WordNet.
- (c) Culture specific considerations in the WordNet.
- (d) Specialities of Indian verbs.

It was emphasised that the glosses in the WordNet explicate the synset senses, but cannot really be encyclopedic, scientific or legal definitions. In explicating the senses, they are assisted by the members of the synset and also the accompanying example sentences. Since the gloss is used for linking and creating the synsets it was decided that

- (a) the glosses will be short and simple.
- (b) they will be expressed both in the specific language and in English.
- (c) the example sentences also will be simple and precise; idiomatic and poetic expressions will be avoided.

The different approaches for building WordNets were discussed. The top ontological categories are "things", "events", "abstracts" and "relationals" which correspond to "concrete nouns", "verbs", "adjectives and abstract nouns" and "postpositions and case markers" respectively. The details of the ontological categories at various levels were discussed. The next was componential approach to the WordNet creation in which, the interesting notion that "words are bundles of semantic features which are binary and parallel to those in phonetics", was introduced. In this approach a space of semantic features is designed and the words sharing ALL and ONLY a set of common semantic features is inserted into the same synset.

Another important issue discussed was the "verbs". Verbs in Indian languages show some unique features like (i) conjunct verbs (ii) compound verbs (iii) causative formation (iv) pairings and (v) onomaetopia.

The computational issue of storing the verbs in the WordNet was also discussed. It was decided that

- (a) Conjunct verbs will be lexicalised in the WordNet.
- (b) Compounds and all the other phenomena will be dealt with by a separate morphological module serving as the front end to the WordNet.

Culture specific concepts were discussed in details. Culture terms do not have their parallels in other languages. Their linkages with the synsets of other languages remains a question? A viable approach could be linking indirectly through the hyperonymy relation. The concepts could be classified as (i)universals across world's languages (ii)universals across Indian languages and (iii)those specific to individual languages. Initially the groups should concentrate on (i)and link with one another.

At the end of the workshop the following resolutions were adopted:

1. By the end of 2003 each Indian language will create a WordNet of 5000 synsets. These will be for about 2000 most frequent content words in each language. Use will be made of the wordlist sorted by frequency- available with the CIIL.
2. The language specific WordNets {are being}/{will be} developed by the following institutions-
 - CIIL, Mysore:** Kannada, Kashmiri, Punjabi, Urdu, Himachali, Malayalam.
 - IIT Bombay:** Hindi, Marathi and Konkani (in collaboration with the Goa research group for the last mentioned)
 - AU-KBC Chennai and Tamil University Tanjavur:** Tamil and Malayalam
 - University of Hyderabad:** Telegu
 - IIT Kharagpur:** Bengali
 - University of Baroda:** Gujarati
 - Utkal University Bhubaneswar:** OriyaReserach groups have to be identified for building the WordNets of Assamese, Nepali and Languages of the North East.
3. As soon as basic 2000 synsets are ready in any Indian Language, CIIL will launch a Website for Indo-WordNet.

Sri. B.D. Jayaram

Central Institute of Indian Languages
Manasagangotri, Mysore – 570 006, India
jayaram@ciil.stpmv.soft.net

Dr. Pushpak Bhattacharyya

Department of Computer Science and Engineering
Indian Institute of Technology
Mumbai- 400 076, India
pb@cse.iitb.ac.in