

### 4.3 WordNet

#### 4.3.1

- ▶ **Name of the Technology :** *Hindi WordNet*
- ▶ **Nature of Technology :** Knowledge resources
- ▶ **Level: (Product / Technology / Sub-system):** Technology
- ▶ **Technical Description of the Technology/ Product including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product:** The Hindi wordnet is an on-line lexical database. The design closely follows the English wordnet. The synonym set {Gar, gaRh} and {Gar, pirvaar}- for example- can serve as an unambiguous differentiator of the two meanings of {Gar}. Such *synsets* are the basic entities in the wordnet. Each sense of a word is mapped to a separate synset in the wordnet. All word sense nodes are linked by a variety of semantic relationships, such as *is-a* (hypernymy/hyponymy) and *part-of* (meronymy/holonymy).  
The lexical data entered by linguists and lexicographers are stored in a MySQL database. The web interface (www.cfilt.iitb.ac.in) gives the facility to query the data entered and browse the semantic relations for the search word. The interface also provides links to extract semantic relations that exist in the database for the search word. So far approximately 10000 synsets have been entered. This corresponds to about 30,000 words in Hindi. There is a morphological processing module as a front end to the system.
- ▶ **Representative Snapshot / screenshot of the Technology / Product:**



- ▶ **Scalability / Portability / Expandability:** Scalable, Portable and Expandable
- ▶ **Readiness of Transfer of Technology (ToT):** Ready in 2 months
- ▶ **Availability of documentation:** Documentations are available
- ▶ **Testing of the Product / Technology:** All technologies are being submitted to STQC
- ▶ **IPR / Open-source :** Open source, subject to permission from MICT
- ▶ **Potential beneficiaries:** The potential beneficiaries are Indian language content creators, interface designers, search engine industry, translators and translation industry, text-mining industry, linguists and typists

- ▶ **User-agency tie-up** : IISC Bangalore, IIT Kharagpur, CIIL Mysore, IIIT Hyderabad; all sister RCs subject to the approval from MICT.
- ▶ **Name and address of the Resource Persons:**  
*Prof. Pushpak Bhattacharyya and  
 Prof. Rushikesh Joshi*  
*Department of Computer Science and Engineering  
 Indian Institute of Technology  
 Mumbai- 400 076 India.  
 E-mal : {pb, rkj}@cse.iitb.ac.in  
 Web : <http://www.cfilt.iitb.ac.in>*

#### 4.3.2

- ▶ **Name of the Technology:** *ORIYA WORDNET (OriNet)*
- ▶ **Nature of the Technology:** Research and Application oriented
- ▶ **Level: (Product / Technology / Sub-system):** Technology

**4. Technical Description of the Technology / Product including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product:** The OriNet system is designed on the basis of multiple lexical database and tools under one consistent functional interface in order to facilitate systems requiring syntactic, semantic and lexical information of Oriya language. The system is divided into two independent modules. One module is developed to write the source files containing the basic lexical data and these files are taken as the input for OriNet system. Lexicographer takes care the major work of this module. Second module is a set of programs by which it accepts the source files, processes it to display for the user and also provides different interface to use other applications. System has been designed using Object-Oriented paradigm according to Oriya language structure with over 1100 lexical entries, which allow flexibility, reusability and extensibility. It also provides X-windows interface to access the data from the OriNet database as per user's requirement. It helps as a lexical resource for Oriya learners and also for expert scholars who are involved in the research in NLP. The system also consists of Oriya Morphological Analyzer (OMA), which takes care of any type of word such as root word or derived word and also provides syntactic information of the word. Presently, more and more lexical entries in the source file are being added and different applicable programs for use in wider are in the process of development range. This S/W is designed using Java and Java Swing (Front-end) for both the Windows-98/2000/NT and the Linux O/S.

- ▶ **Representative Snapshot / screenshot of the Technology / Product:**



- ▶ **Scalability / Portability / Expandability:** All
- ▶ **Readiness of Transfer of Technology (ToT):** Yes
- ▶ **Availability of documentation :** Yes
- ▶ **Testing of the Product / Technology :** Under progress
- ▶ **IPR / Open-source:** IPR (No- SW-1178-2003)
- ▶ **Potential beneficiaries: Common man & Research Scholar**
- ▶ **User-agency tie-up :** Solicited
- ▶ **Name and address of the Resource Person:**  
*Dr (Ms) Sanghamitra Mohanty  
 RC-ILTS-Oriya, Dept. of CSA,  
 Utkal University, Bhubaneswar, Orissa.  
 E-mail : sandhami@rediffmail.com  
 Web : http://www.its.utkal.org*

### 4.3.3

- ▶ **Name of the Technology:** *Marathi WordNet*
- ▶ **Nature of Technology:** Knowledge resources
- ▶ **Level: (Product / Technology / Sub-system):** Technology
- ▶ **Technical Description of the Technology/ Product including Basic block diagram, Algorithm used, O/S used, Front-end / user interface, and Specification of the Technology / Product :** Since Marathi and Hindi share many common features, it has been possible to adopt the Hindi wordnet as the basis for Marathi wordnet creation. Hindi and Marathi originate from *Sanskrit* and, therefore, the  $t\%sama$  (words taken directly from mother language, *e.g.* gait, kRit) and the  $td\Bava$  (a word which has evolved organically from the mother language, *e.g.* gaRh à Gar, kma- à kama) words in both the languages are often same in meaning. Also, both the languages have the same script- *Devanagari*.

In the construction of the Marathi wordnet, a Hindi synset is chosen and mapped to the corresponding Marathi synset. For instance, for the Hindi synset {**kagad, kagaja**}- standing for paper- the Marathi synset is {**kagad**}.

So far, 3400 synsets have been entered. This corresponds to about 10,000 words in Marathi. Our aim is to cover all the common Marathi words. As the Marathi wordnet is based on the Hindi wordnet, all the semantic relations become automatically inherited. An additional benefit to accrue is the creation of a parallel corpus.

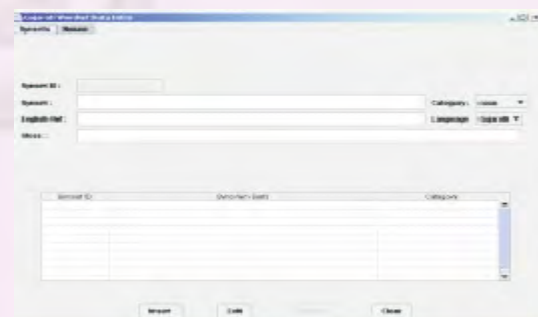
- ▶ **Representative Snapshot / screenshot of the Technology / Product:** Same as Hindi-Word-Net
- ▶ **Scalability / Portability / Expandability:** Scalable, Portable and Expandable
- ▶ **Readiness of Transfer of Technology (ToT):** Ready in 2 months
- ▶ **Availability of documentation:** Documentations are available

- ▶ **Testing of the Product / Technology:** Product submitted to STQC for testing
- ▶ **IPR / Open-source:** Open source, subject to permission from MICT
- ▶ **Potential beneficiaries:** The potential beneficiaries are Indian language content creators, interface designers, search engine industry, translators and translation industry, text-mining industry, linguists and typists
- ▶ **User-agency tie-up :** IISC Bangalore, IIT Kharagpur, CIIL Mysore, IIIT Hyderabad; all sister RCs subject to the approval from MICT.
- ▶ **Name and address of the Resource Persons:**  
*Prof. Pushpak Bhattacharyya and  
 Prof. Rushikesh Joshi  
 Department of Computer Science and Engineering  
 Indian Institute of Technology  
 Mumbai- 400 076 India. {pb, rkj}@cse.iitb.ac.in  
 Web : <http://www.cfilt.iitb.ac.in>*

#### 4.3.4

- ▶ **Name of the Technology :** Gujarati Word-Net (Lexical Resource)
- ▶ **Nature of the Technology :** Knowledge Resource
- ▶ **Level :** Product may be developed from the technology
- ▶ **Technical Description :** Gujarati WordNet, a Lexical Resource, is in the process of development. RC at MSU is working in close collaboration with and through able guidance of Professor Pushpak Bhattacharya of RC at IIT, Mumbai. The Data Entry Interface for entering the synsets is build. Presently above 500 synsets are already been developed. Further work on preparing the synsets and entering into the database is going on. Basic Semantic Relationships are also been established like Hypernymy, Hyponymy between the synsets formed and entered in the database. Further work is going on for establishing other Semantic relationships between the synsets. The Web Interface is already under construction to facilitate the access of an On-Line Lexical Database of Gujarati to be available over the Internet.

#### ▶ Representative Screenshot



- ▶ **Operating System used :** Windows 98, 2000, XP
- ▶ **Scalability / Portability / Expandability :** Available
- ▶ **Availability of documentation :** Yes
- ▶ **Testing of the Product / Technology :** a Testing has been done
- ▶ **IPR / Open Source :** IPR lies with MS University of Baroda & Dept of Information Technology
- ▶ **Potential beneficiaries :** Indian language content creators, search engine industry translators, text-mining industry linguists, typists.
- ▶ **User agency tie-up :** Solicited
- ▶ **Name and address of the Resource Person:**  
 Prof. Sitanshu Mehta  
 M.S.U. Baroda, Vadodara, Gujrat