

# Preface

## IndoWordNet: A Rainbow in the Indian Lexical Panorama

### 1. The Mission

The IndoWordNet is a consortium-mode multilingual project that was to develop WordNets for seven Indian languages, namely (in alphabetical order) Bangla, Gujarati, Kashmiri, Konkani, Odia, Punjabi, and Urdu. It is funded by the Department of Electronics Information Technology (DeitY), Ministry of Communication & Information Technology (MCIT), Govt. of India, and is executed by a consortium of nine academic institutions of India, namely the Goa University, Goa (the Consortium Leader); Indian Institute of Technology Bombay, Mumbai; Indian Statistical Institute, Kolkata; Dharmasinh Desai University, Nadiad; University of Kashmir, Srinagar; University of Hyderabad, Hyderabad; Punjabi University, Patiala; Thapar University, Patiala; and Jawaharlal Nehru University, New Delhi. The final deliverable of the project is the integrated WordNet (<http://www.cfilt.iitb.ac.in/indowordnet>) consisting of minimum of 30,000 linked synsets for each of the seven languages included in the project (Table 1).

The work on development of WordNet for Indian languages started in 2000 when the Natural Languages Processing group from the Center for the Indian Language Technology, Department of Computer Science and Engineering of the Indian Institute of Technology, Bombay (IIT-B), initiated an effort for developing the Hindi WordNet (<http://www.cfilt.iitb.ac.in>). This digital resource was made publicly available in 2006 under the GNU (<http://www.cfilt.iitb.ac.in/wordnet/webhwn>). It was generated with the financial support from the Technology Development for the Indian Languages (TDIL) (<http://tdil.mit.gov.in/>) project of the Ministry of Communication and Information Technology (MCIT), Govt. of India, and the Ministry of Human Resources Development (MHRD), Govt. of India. It was the first WordNet among the Indian languages, and the methods and strategies adopted for developing this WordNet was almost the same as that of the Princeton WordNet for English. (<http://wordnet.princeton.edu/>). The WordNet for other Indian languages (e.g., Indradhanush WordNet and Dravidian

**Table 1** Synset generation statistics of IndoWordNet Consortium as on June 2015

S.N.	Name of Institute	Nouns	Adjectives	Verbs	Adverbs	Total
1.	Goa University (Konkani)	22,985	5644	2983	450	32,062
2.	Indian Institute of Technology Bombay (Hindi)	28,312	6090	3108	461	37,971
3.	University of Kashmir (Kashmiri)	21,041	5365	2660	400	29,466
4.	University of Hyderabad (Odia)	27,216	5273	2418	377	35,284
5.	Punjabi University and Thapar University (Punjabi)	21,573	5830	2836	443	30,682
6.	Jawaharlal Nehru University (Urdu)	21,595	5787	2800	443	30,625
7.	Indian Statistical Institute (Bangla)	27,281	5815	2804	445	36,345
8.	Dharmsinh Desai University (Gujarati)	24,896	5828	2805	445	33,974

WordNet) followed the same course of action initiated for the Hindi WordNet. A large nationwide project for building WordNets for the Indian languages was conceptualized as the IndoWordNet, which was conceived under the active guidance obtained from the WordNet Team of the IIT-B, Mumbai. The IndoWordNet, in principle, is a linked lexical knowledgebase of WordNets of 18 scheduled languages of India (in alphabetical order): Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Odia, Punjabi, Sanskrit, Tamil, Telugu, and Urdu. The Indradhanush WordNet Consortium is one of the core members of the IndoWordNet Project.

## 2. The Learning Experience

Every research work is a journey from the known to unknown through the terrains of trials and errors with a sanguine scope for learning and knowledge gathering. Our Indradhanush project is no exception. The experience which we have gathered and the knowledge which we have gained from this consortium project may be summed up as follows.

While creating culture-specific synsets (CSS) for each member language of the WordNet, we have come across many unfamiliar concepts and ideas, which are not only unique to a particular language or a culture, but are also instrumental in the generation of new information on and insights into the multilayered fabric of human life and living. For instance, the term '*bhairavjap*' in Gujarati refers to the act of embracing death by jumping from the top of a high mountain—as a part of self-sacrificing *yajna* or prayer to the almighty which is practised by a specific religious group. Such a unique practice of the specific religious community of Gujarat is rarely known to other speech and cultural communities in India.

The border languages or adjacent languages share many cultural concepts and ideas, which are the outcomes of close cultural interactions between the communities due to their geographical proximities. For instance, many concepts in Konkani,

which are also found in Marathi and Kannada, are hardly observed in other language communities. Similar situations are observed for Bangla and Odia, Kashmiri and Punjabi, or Bodo and Assamese, even though some of these languages are not always genealogically linked. It would be right in assuming the existence of such cultural meeting points between Odia and Telugu due to their geographical closeness, even though they belong to two different language families of India.

While working on CSS, we realized that there is an urgent need of pictorial representation of items and objects that are claimed to be unique to a particular culture or community. Although it is not always possible to give a picture or an image for an abstract idea or an elaborate function carried out by the members of a culture or language, it is always sensible to have images and pictures for concrete items and objects, so that people of other languages and cultures may either identify similar items and objects in their culture to record cultural and linguistic solidarity or create the new concepts properly. For instance, in Konkani, we come across a new term ‘pilot,’ which refers to a ‘machine-pulled rickshaw that is normally used as a carrier of goods and commodities in the state.’ Although it seemed to be a unique idea at the initial stage, the actual picture of the vehicle helped others to confirm that it is not a new concept, since a similar type of vehicle is also found in other cultures and language communities. For instance, look at the pictures given below:



Since it is presumed that the present WordNets will not only serve the online lexical queries of end users, but also be used for machine translation across Indian languages (as well as between English and Indian languages), it is rational that CSS should have their legitimate places in the synset database of the member languages so that cross-lingual translation and information sharing are not blocked. For instance, the concept of '*bihu*,' which is a unique cultural event of Assamese speech community, may be incorporated in the CSS of Assamese with elaborate audio-visual information and reference so that the people of other language communities can access the concept, and, if required, are able to translate (manually or automatically) the concept in their respective languages in an appropriate manner. Similar treatment may be extended to the concept of '*baishakhi*' in Punjabi, '*pongal*' in Tamil, '*onam*' in Malayalam, '*raja*' in Odia, or '*bhaadu*' in Bangla.

Sense marking of words in corpus of Indian languages does help increase the coverage of the concepts in WordNets as well as enhance their quality and utility potential of the resource. It is indeed practiced on a trial basis to see whether the senses that are recorded and presented against each synset listed in the WordNet are at all inclusive of the senses that are found to be assigned to the terms when these are actually used in language texts. Some preliminary surveys have shown that many of the senses of the synsets are not captured in the WordNets. Conversely, many of the senses that are stored in a WordNet are not always found in the corpus texts. This implies that the knowledge base as well as the senses assigned to the synsets needs to be verified, validated, and updated against the data and information retrieved from the corpora of Indian language texts.

While marking senses of words in newspaper texts, we find that newspaper texts of corpora of present-day Indian languages contain many named entities, transliterated words, item names, function words, abbreviations, acrostic words, etc., which are not incorporated in the WordNets. This raises a serious theoretical question in regard to the list of synsets included in the WordNet: Should we keep only universally approved (i.e., present across all languages) nouns, verbs, adjectives, and adverbs or should we include rare words in all parts-of-speech as well as terms of specialized domains, areas, and fields? Also the question of including non-Indian proper names and foreign names as well as non-Indian words in the synset list is a crucial issue that deserves serious attention and careful deliberation.

The making of Indradhanush WordNet is actually a long learning process of discovering India through the meanders of language diversity as well as understanding India through the terrains of cultural difference. Even if there are registered and acknowledged diversions in Indian languages and culture, the Indradhanush WordNet has taught us to imbibe the skill to bring all Indian languages under one umbrella and accommodate all conceptual diversities within a single frame of a synset. Moreover, it has taught us to develop unique ways to work in a consortium environment, where multiple teams from different institutes across the nation can work in an integrated manner to develop the resources and solve the challenges. The teleconferences among the member teams helped us realize the impact of collective team work with shared responsibilities—a unique feeling that incorporated solidarity among the members of a collective effort.

Finally, the creation of WordNet resource and WordNet home pages has given us ample opportunity and valuable experience to be creative while working. We had to be scientific in approach and artistic in attitude; informative in data and innovative in explanation; and analytic in interpretation and methodical in presentation. We had to club together both information and imagination in an innovative manner so that a lifeless and monochromatic concept becomes lively and colorful in the audio–video format of the home page.

### 3. Development of Tools and Utilities

Member teams of the Indradhanush WordNet have worked together to identify the problems relating to the development of the system and the resource. In the process of addressing these problems, some of the more experienced and expert members have been assigned the tasks for developing required tools and other devices:

- (a) Synset Categorization Tool is developed by IIT-Bombay to choose common linkable synsets across all languages by classifying them as Universal, Pan-Indian, Language Specific, etc.
- (b) Synset Creation Tool is developed by IIT-Bombay. It is an offline interface to create target language synsets by using Hindi language synsets as source data.
- (c) Sense Marker Tool is developed by IIT-Bombay to track the amount of synset coverage of a WordNet in the newspaper corpus of respective Indian languages.
- (d) Generic Stemmer for Indian Languages is developed by IIT-Bombay to find out possible stems of a given inflected/suffixed word used in newspaper corpus.
- (e) WordNet Linkage Tool is developed by IIT-Bombay to link up Hindi WordNet with the English WordNet. The tool uses 13 different heuristics to automatically identify top 5 English synsets for a given Hindi synset.
- (f) Word Sense Disambiguation Tool is developed by IIT-Bombay. It provides single access point to 9 different state-of-the-art word sense disambiguation algorithms.
- (g) WordNet Content Management System (CMS) [v1.0, v2.0] is developed by Goa University. It allows creation of WordNet Web sites with versatile user interface and desired functionalities in a very short time for many Indian languages.
- (h) CSS Manger Tool (v1.0) is developed by Goa University. It is a centralized Web-based tool that assists in creation of Concept-Specific Synsets (CSS) as well as manages their linkages to the WordNet of other Indian Languages. It can also be used for validation of synsets in the WordNets.
- (i) Sense Marking Statistic Finder Utility Tool is developed by Goa University. It assists to find coverage statistics of the sense-marked corpus.
- (j) Synset Merger Utility is developed by Goa University. It merges different synset files into one single file for synchronized access.

### 4. Development of Web Sites and Digital Resources

One of the primary goals of the project was to design and develop separate Web sites for the WordNet of each of the languages included in the Indradhanush project. It was a part of the primary deliverables of the project which could be directly accessed by the users as an open source resource for Indian people—the mission for which the TDIL and the DeitY, as well as the MCIT, Govt. of India, are committed

to. It is a happy moment for the entire consortium to declare with some satisfaction that the all language groups have successfully developed separate WordNet home pages which are operational and globally accessible. The IP addresses of each of the WordNet Web sites are given below for general reference and access.

- Indradhanush WordNet Consortium: <http://indradhanush.unigoa.ac.in/>
- Bangla WordNet: <http://www.isical.ac.in/~lru/wordnetnew/>
- Gujarati WordNet: <http://www.cfilt.iitb.ac.in/gujarati/>
- Kashmiri WordNet: <http://indradhanush.unigoa.ac.in/kashmiriwordnet/>
- Konkani WordNet: <http://konkaniwordnet.unigoa.ac.in/>
- Odia WordNet: <http://indradhanush.unigoa.ac.in/odiawordnet/>
- Punjabi WordNet: <http://punjabiwordnet.com/>
- Urdu WordNet: <http://indradhanush.unigoa.ac.in/urduwordnet/>
- IndoWordNet Database v1.0, v2.0, v3.0—by Goa University: Relational database structure to store WordNet data and relationships.
- <http://indradhanush.unigoa.ac.in/public/downloadTools/downloadTools.php>
- IndoWordNet API—v1.0, v2.0, v3.0—by Goa University: IndoWordNet Application Programming Interface (IWAPI) helps in providing access to the WordNet resources independent of the underlying storage technology.
- <http://indradhanush.unigoa.ac.in/public/downloadTools/downloadTools>
- <http://www.tdil-dc.in/indowordnet/>

A close look into each of the WordNet Web sites clearly shows how life, culture, custom, and history of each language is distinctly reflected in the Web site even though each Web site is developed following the same structure and composition of the Hindi WordNet Webpage. Another vital component of these Web sites is the option of feedback from the WordNet users—based on which each Webpage desires to update its content, upgrade its composition, and improve its application. In fact, future growth and application relevance of these Web sites heavily depend on how the end users utilize these resources and how these online portals are able to serve various linguistic and extralinguistic requirements of the people of the country.

## 5. Milestones Achieved

Several milestones have been reached in the course of the project:

- We have developed one of the largest lexical databases in digital form for seven Indian languages. The uniqueness of the lexical stock is that it is not just a compilation of lexical items in some order or other; rather, it is repository where lexical items are systematically classified based on certain predefined parameters, such as part-of-speech and lexical type.
- The lexical resource is the first of its kind in Indian languages. It is a useful free lexical resource available to general people. It is a great learning experience for a learner to find conceptually equivalent words—for a synset of his/her query—from seven Indian languages.
- It is an important part for building machine translation system across Indian languages and English. By ensuring the interlinking of synsets among the Indian languages and also with English, we can provide the basic means of translation of words.



- People who want to learn a new Indian language can use this resource effectively access logs and feedbacks bear testimony to this.
- The resource is enabler of word sense disambiguation (supervised and unsupervised) research in Indian languages. This leads to the development of state-of-the-art WSD mechanisms leading to international publications.
- It has been possible to develop standard sense-marked (annotated) corpora in Indian newspaper texts, which are useful both for WordNet and for semantic role labeling, for example, in Universal Networking Language (UNL).
- Cross-lingual Information Retrieval (CLIR) systems use linked WordNets query translation.
- Similarity computation has been possible due to WordNets.
- The bilingual and multilingual dictionaries that are in the state of compilation from the WordNet are useful for manual as well as statistical machine translation.
- Urdu is the official language of Jammu and Kashmir. Even then, there is no online Kashmiri–Urdu dictionary. The present WordNet may be used as an online Kashmiri–Urdu dictionary as well. Furthermore, it can be used as Kashmiri–English dictionary with more than thirty thousand entries.
- In fact due to the linked nature of WordNets, it has been possible to create bilingual mappings for 18 languages ( $18 \times 17$  pairs) which are freely downloadable for research purposes (<http://www.cfilt.iitb.ac.in/Downloads.html>).
- This project has succeeded in establishing a unique synergy between computer scientists and linguists. This cross-fertilization of disciplines, on one hand, has helped linguists understand the nature and complexities of computation, and on the other hand, has helped computer scientists, realizes the intricacies of the way natural languages operate.

## 6. Technological Spin-Offs

IndoWordNet has emerged as a digital lexical knowledge base of 18 different Indian languages. This resource is used in many NLP projects related to Indian languages, such as ILILMT (Indian Language to Indian Language Machine Translation), CLIA (Cross-lingual Information Access), and Indian language sentiment analysis.

Multilingualism presents a major challenge for developing a semantic Web in a multilingual country such as India, where information spread across Indian languages needs utmost care for its processing and synchronization. Since millions of people like to access relevant content in their own native tongues, it is required to form a framework or an interface which can provide scopes for information sharing and reuse across the Indian languages. In order to achieve this goal, it is important to make WordNet resources available across the Indian languages through a common representation format so that people can share data across the languages.

An offshoot of the IndoWordNet project is the semantic Web porting of Indian language WordNets. The worldwide semantic Web project aims to achieve complete interoperability and linkage of Web data (linguistic linked open data or LLOD). The steps toward this aim are (i) converting WordNets of different Indian languages into RDF format and (ii) creating resources and tools that can be used

to develop semantic Web application for Indian languages. Major stages and outcomes of the project are as follows:

- Conversion of the existing IndoWordNet database to RDF (resource description language) format.
- Development of mechanism to use IndoWordNet for Multilingual semantic search.
- Development of IndoWordNet applications which can help larger population in various tasks such as computer-assisted language learning and multilingual dictionary.
- Enrichment of IndoWordNet through gamification.
- Cognitive Study of lexical and relational semantics using eye-tracking mechanism.

## 7. The Present Volume

This volume contains 11 research papers presented in the 2nd National Workshop of Indradhanush WordNet held at Goa University, Goa, from August 8 to August 10, 2011. Also, it includes two papers related to Tamil and Malayalam WordNets—two major components of the Dravidian WordNet which in turn is a part of the IndoWordNet.

While some papers raise theoretical questions in regard to concept definition and encapsulation in the form of WordNet synsets, majority of papers provide details of the problems faced by the researchers during the process of generating conceptually equivalent synsets in their respective languages. The problems encountered during finding equivalent of Hindi synsets and example sentences in their respective languages are discussed by quite a few authors.

In majority of the cases, the authors have not only identified the problem areas which demand serious investigation into the texture and information embedded within a synset and its elaboration, but they have also have provided some solutions to overcome these problems.

We visualize all those interested in WordNets of Indian languages to be the target users of this volume. The included papers are neither highly technical in approach, nor rigorously critical in data analysis, nor elusively intellectual in presentation of content. The book, therefore, may be used as a reference book on Indian language WordNets by one and all. It can be specifically useful for those who are working in WSD in Indian languages. It can be used by teachers who use WordNets to teach Indian languages. Moreover, language researchers, linguists, grammarians, morphologists, and language technologists, who are working in different areas of descriptive, applied, and computational linguistics, can benefit from this volume. In essence, the anthology is a valuable reading for NLP scholars, linguists, lexicographers, and language teachers.

We shall consider our efforts successful if people find this volume useful for their respective needs.

Kolkata, India  
Mumbai, India  
Taleigao, India

Niladri Sekhar Dash  
Pushpak Bhattacharyya  
Jyoti D. Pawar



The WordNet in Indian Languages

Dash, N.S.; Bhattacharyya, P.; Pawar, J.D. (Eds.)

2017, XVII, 264 p. 76 illus., 59 illus. in color., Hardcover

ISBN: 978-981-10-1907-4