

Chapter 2

Insights on Hindi WordNet Coming from the IndoWordNet

Laxmi Kashyap, Salil Rajeev Joshi and Pushpak Bhattacharyya

Abstract In a multilingual country such as India, machine translation and crosslingual search are highly relevant problems. The WordNets, as crucial linguistic resources, play the most dominant role in the field of text processing and applications, such as machine learning, machine translation, information extraction, information retrieval, and natural language understanding systems. Therefore, no meaningful research in these areas can be complete without their help. This paper reports the categorization work of synsets of the Hindi WordNet (version 1.2), the challenges that were faced while doing the work, and solutions obtained for them thereafter. There are a number of concepts common to most of the languages, and linking these concepts with each other can provide an indispensable resource for Natural Language Processing and Language technology. The WordNet for Hindi language is created using the ab initio method while all the other Indian language WordNets are being created using the Hindi WordNet through expansion approach. The Hindi WordNet forms the foundation for the other Indian language WordNets as they are based on it and are being linked to it.

Keywords Language-specific synset • Hindi WordNet • IndoWordNet • In-family synset

L. Kashyap (✉) · S.R. Joshi · P. Bhattacharyya
Department of Computer Science and Engineering,
Indian Institute of Technology-Bombay, Powai, Mumbai, India
e-mail: yupu@cse.iitb.ac.in

S.R. Joshi
e-mail: salilj@cse.iitb.ac.in

P. Bhattacharyya
e-mail: pb@cse.iitb.ac.in

2.1 Introduction

Among the Indian language WordNets, the Hindi WordNet (HWN) was the first one to come into existence from 2000 onward. It was inspired by the English WordNet which contains nouns, verbs, adjectives, and adverbs organized into synonym sets, each representing one underlying lexical concept. Different relations such as hypernymy and hyponymy link the synonym sets to each other. Soon, other Indian language WordNets started getting created. The WordNets for Marathi and Sanskrit followed the Hindi WordNet. All these three efforts are going on at Indian Institute of Technology, Bombay. Other Indian language WordNets are being linked to Hindi WordNet, paying particular attention to language-specific phenomena. Thus, linking Hindi WordNet to the English WordNet and then linking other Indian language WordNets to Hindi, in turn, will help to increase the linkage of concepts and will create a wide WordNet grid of shared concepts.

The Indradhanush project is formed to develop WordNets of Indian languages, which currently comprises seven major languages across India, viz., *Bengali*, *Gujarati*, *Kashmiri*, *Konkani*, *Panjabi*, *Oriya*, and *Urdu*. The construction of individual WordNet is carried out mostly by different Indian organizations. The **Bangla WordNet** is being created at the *Indian Statistical Institute*, Kolkata; the **Gujarati WordNet** is being created at the *DD University*, Nadiad; the **Kashmiri WordNet** is being formed at the *University of Kashmir*, Srinagar; the **Konkani WordNet** is being developed at the *Goa University*, Goa; the **Panjabi WordNet** is being generated at the *Thapar University*, Patiala; the **Odia WordNet** is being developed at the *Central University*, Hyderabad; and the **Urdu WordNet** is being developed at the *Jawaharlal Nehru University*, New Delhi.

The road map of the paper is as follows: Sect. 2.2 presents the creation of Hindi WordNet and its principles. Section 2.3 deals with the concept of categorization and relevant information. Section 2.4 describes the challenges involved in the task as well as the specific fields where they occur. Section 2.5 presents the solutions. Section 2.6 describes related tools and contains their snapshots. Section 2.7 presents the linkage statistics, and finally, Sect. 2.8 winds up discussion with conclusion and future work.

2.2 Creation and Principles of Hindi WordNet

Hindi WordNet creation started in the year 2000 with the help of a specially designed data entry tool. A common word from a Hindi dictionary is chosen. Then, the particular concept is made explicit with the help of elaborate gloss and example(s). Other synonyms according to that concept are taken and then definition, example, and ontology are added to the word. The part-of-speech is selected and, based on this, relations of the words are established.

To create a Hindi synset, three principles are adopted. They are as follows: minimality, coverage, and replaceability. However, the principle **minimality** implies that minimal sets of words are chosen to make the concept unique. For instance, both ‘ghara’ and ‘griha’ constitute a minimal set of words which denotes the concept of a ‘house.’

The principle **coverage** implies that the maximal set of all possible words that stands for the sense and ordered by frequency in the corpus are included in the list of synonymous words. For instance, ‘makaana,’ ‘sadana,’ ‘shaalaa,’ ‘aalaya,’ ‘dhaama,’ ‘niketana,’ ‘vaastu,’ and ‘paNa’ are the synonymous words for the concept of ‘house.’

Finally, the principle **replaceability** refers that the example sentences are such that the most frequent words in the synset can replace one another in the sentence without altering the sense, for instance, *isa ghara men paancha kamare hain* ‘There are five rooms in this house.’

2.3 Categorization

2.3.1 Need for Categorization

When the IndoWordNet project started, there were about 32000 synsets in Hindi WordNet. The question arose about the set of synsets which were to be sent to language groups in which WordNets for respective languages were to be created as a part of this project. The reason for this activity was that the Hindi synsets were created alphabetically and not according to their most common uses. Since there are many uncommon synsets in the first five thousand synsets which do not have any lexeme in many of the languages falling under the project, it was difficult to send IDs serially. Hence, it was decided to categorize them according to their uses in different categories.

Synset categorization is an act of distributing synsets into categories of different kinds. This was done so that other WordNets can make the most commonly used synsets first and then go on to other categories such as (a) **Core**, (b) **Common**, (c) **Common in Indian language**, (d) **Common in Hindi**, and (e) **Uncommon**; the original categories are chosen for the purpose. This categorization was done at the IIT Bombay in March before the first IndoWordNet workshop was organized in Coimbatore from June 11 to June 14, 2009.

2.3.2 Original Categorization

In the original categorization, only **core** and **common synsets** were given more importance, since those synsets had to be linked in the first place. The **core synsets**, which are necessary for the day-to-day communication, were selected from

Bhaaratiiya Vyavahaara Kosha compiled by D. N. Naravane (1961). It is a multilingual dictionary which has lexical terms in sixteen (16) Indian languages for a particular concept. Nearly 2,000 concepts were selected as the core synsets from this book. These synsets were separated from the HWN senses.

After the core synsets, attention was devoted to **common synsets**. The common concepts were selected from remaining 32,000 synsets. The synsets were distributed among six (6) people to rank them into following categories with the help of a specially designed ranking tool (description and snapshot in Sect. 2.5): Common, Common in Indian languages, Common in Hindi, and Uncommon. In this way, 16,000 synsets were categorized as common synsets. These were again confirmed by manual voting and 12,000 synsets were finally selected in this process as common synsets.

2.3.3 Categorized Synsets to Be Linked to Other Indian Languages

The core and common synsets were the first lot of the synsets sent to the IndoWordNet group along with the off-line tool (description and snapshot in Sect. 2.6) to be linked with other Indian languages. IndoWordNet consists of *Assamese, Bengali, Bodo, Gujarati, Kannada, Kashmiri, Konkani, Manipuri, Malayalam, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu*, and *Urdu* language groups (Bhattacharyya 2010).

2.3.4 Discrepancies in Categorization

The IndoWordNet language groups met at the Shillong Symposium held on April 12 to April 14, 2010, and discussed the synset categorization and raised the concerns regarding discrepancies in synsets and their categories. According to the opinion of different language groups, some synsets which were categorized as core synsets were actually belonged to common synsets category, and vice versa. Followings examples were categorized as core synsets while they should have been in the common category.

- *anuvaada*—translation (Noun)
- *aparaadhii*—criminal (Adjective)
- *aastika*—theist (Noun)
- *daana*—dotation (Noun)

In the reverse manner, the followings were categorized as common synset, while they should have been in the core category.

- *sheranii*—lioness (Noun)
- *saarangii*—a string musical instrument (Noun)
- *godhulii Belaa*—dusk (Noun)
- *karelaa*—bitter gourd (Noun)

2.3.5 New Categorization

The exercise of categorization was repeated to resolve the problems stated above. As a result, new categorization was done in May 2010 and this time synsets were categorized into six (6) categories instead of the initial four (4) categories by different WordNet groups.

The first group consisted of Gujarati and Konkani groups to whom synsets from 1 to 11,400 were given; Nepali and Punjabi groups were in second group to whom synsets from 11,401 to 22,800 were given, and third group included Telugu and IIT-B groups to whom the rest of the synsets in Hindi WordNet were given. Following are the new six (6) categories in which the synsets are divided:

- (a) **Universal Synset:** Synsets which have an indigenous lexeme in all the languages of the world (e.g., *sun*, *earth*).
- (b) **Pan-Indian Synset:** Synsets which have indigenous lexeme in all the Indian languages but no English equivalent (e.g., *paapaDa*).
- (c) **In-family Synset:** Synsets which have indigenous lexeme in the particular language family (e.g., *bhatijaa* in Dravidian languages).
- (d) **Language-specific Synset:** Synsets which are unique to a language (e.g., *bihu* in Assamese)
- (e) **Rare Synset:** Synsets which express technical terms (e.g., *ngram*, etc.).
- (f) **Synthesized Synset:** Synsets created in the language due to influence of another language (e.g., *pizza*, etc.).

2.3.6 Challenges in New Categorization

Hindi synsets of some very core concepts did not have their counterparts in English. They, therefore, could not be categorized as Universal since the definition of it says that the concept should have an indigenous lexeme in all the Indian languages as well as in English, e.g., ‘luhaarii’ there is no English word for the Hindi concept of ‘work of blacksmith.’

Hindi synsets having very core concept did not have the required equivalent synset in English WordNet. Therefore, these concepts could not be selected as Universal. For instance, ‘soyaa huaa’ ‘dormant, sleeping,’ but the English WordNet gives only one sense, that of lying with head on paws as if sleeping. This sense cannot be applied to humans; hence, the term cannot be categorized as a core

concept. Similarly, ‘anuvaada’ refers to ‘translation, interlingual rendition, rendering, version (a written communication in a second language having the same meaning as the written communication in a first language.’ Since this sense cannot be applied to oral translation, the term cannot be categorized as a core concept.

Hindi adjectival concepts, for which mostly *-ed* and *-ing* forms of English adjectives are used, are not found in English dictionaries. Therefore, these concepts were also excluded from Universal categories, although, rightfully, they belong to here. For example, ‘pisaa huua’ should be equivalent to ‘milled.’

Many other Hindi words also do not find their matching terms in English lexicon. Therefore, these synsets were also not selected as the Universal synsets, for instance, ‘daaniya’ ‘donatable,’ and the causative verb, such as ‘sulaanaa’ ‘to make someone sleep’.

For many Hindi words that are used as adjectives as well as nouns, only noun forms are given but adjectives are not present in the English WordNet. This was another reason for excluding some synsets from Universal category although it was felt they were the core concepts, for instance, ‘dalabadaloo’ ‘defector’ (noun).

Lack of knowledge whether a synset has an indigenous lexeme in all other Indian languages or in a particular language family was also a hurdle for categorization.

Categorization of synsets between rare and synthesized categories was also not easy because of disagreement among the lexicographers, e.g., ‘pneumonia’ and ‘diphtheria.’

2.3.7 Yet Another Categorization

Sighting the above obstructions, in the workshop held at the Indian Institute of Technology, Kharagpur, on December 8, 2010, it was decided to categorize all the 34,378 Hindi synsets again. These synsets were sent to all the IndoWordNet groups, along with a categorizer tool (description and snapshot in Sect. 2.7), in which they had following options to categorize the synsets:

- (a) Yes : If an indigenous lexeme for the synset exists in the language
- (b) No : If there is no indigenous lexeme for the synset in the language
- (c) Not set : If the concept in Hindi is not clear to the group.

Depending on the responses from all the language groups, the following two categories were populated:

- (a) **Universal synsets:** The synsets which were categorized as ‘Yes’ by all the groups, and also had equivalent English words or synsets, were classified here.
- (b) **Pan-Indian synsets:** The synsets which were categorized as ‘Yes’ by all the groups, but did not have equivalent English words or synsets, were classified under this category.

2.4 Challenges in Linking

The linkage task has to do a fine balance between maintaining accuracy and providing maximum linkages. While linking with Hindi WordNet synsets, other language WordNets creator encountered several challenges. The problems have occurred because the languages belong to different language families. The specific areas where such problems were faced are listed below.

2.4.1 Challenges in Linking Synsets Across Languages

Other language groups had difficulties in linking with the Hindi WordNet synsets due to the following reasons:

2.4.1.1 Kinship Relation

Kinship relations in Hindi such as ‘bhatijaa’ ‘a brother’s son’ could not be linked directly in Dravidian languages since they have many terms to denote this concept, depending on gender of the speaker and whether the referent or speaker is elder or younger to him. In Manipuri, they have three different terms for the Hindi concept ‘betaa’ as a ‘male child,’ viz., ‘nupaamachaa’ for ‘a boy,’ ‘ichaanupaa’ for ‘my son,’ and ‘machaanupaa’ for ‘his son’ which is also a generalized term. The same kind of feature is in Bodo in the case of ‘Maana.’ For instance, ‘bhatijaa’ ‘a brother’s son’ in Hindi has no equivalent lexeme in Telugu, but Telugu has different terms for this concept, a couple of which are mentioned below:

Telugu: Tam’muDu kumaaruDu
Hindi: (*chhote bhaaii kaa betaa*)
Meaning: ‘younger brother’s son’

Telugu: (Annayya kumaaruDu)
Hindi: (*bade bhaaii kaa betaa*)
Meaning: ‘elder brother’s son’

2.4.1.2 Change in Part-of-Speech

The predicative adjectives always become verbs in Manipuri language. Therefore, they cannot be linked as such. For instance, look at the following examples,

Manipuri: Nupi machhaa phajei
Hindi: (*laDakii sundara hai*)
Meaning: ‘The girl is beautiful’

Some Hindi concepts cannot be linked with Sanskrit due to change in their part-of-speech in the language. However, these words are available in instrumental cases of nouns/adjectives in the same sense. For example:

Hindi: (shaantipoorvaka)

Definition: (shaanti ke saatha yaa shaanti se) (adverb)

Meaning: 'quietly or with quietness'

Example: (aapa saba shaantipoorvaka merii baaton ko sunen)

English: 'all of you listen to my speech quietly'

2.4.1.3 Different Terms in a Language for a Hindi Concept

There are two terms for water in Kashmiri language, whereas Hindi has only one term for water. The general term for water is 'aaba' and the term for drinking water is 'tresh.' Therefore, the question arises as to which one should be linked with the Hindi concept.

2.4.1.4 Language-specific Concept

The concepts which are very core to a language, such as its culture and food, are not there in the Hindi WordNet. So the question remains about how these terms can be linked with the Hindi WordNet. For instance, in *Marathi naoovaarii, navavaarii—naoo vaara laamba asalelaa* 'tii naoovaari lugaDe nesate' naoovaarii, navavaarii—which is nine yard long 'She wears naoovaari saari.'

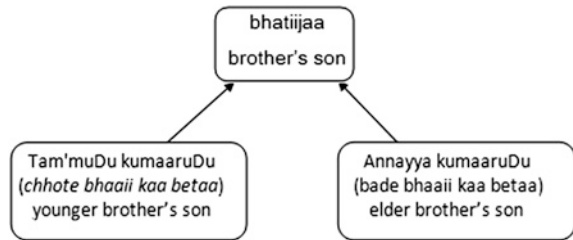
2.4.1.5 Challenges in Linking with English Synsets

Many challenges were encountered while linking Hindi synsets with English synsets. The problems have occurred because the two languages belong to cultures and social mores which are widely different. Musical instruments, kitchen utensils, tools, spices, grain, cast, occupation, wages, women, denoting cast and occupation, food, culture, etc. of Hindi synsets could not be linked to the English WordNet, since English does not have different terms for them (Saraswati et al. 2010), for instance, 'tabalaa,' 'dholaa,' 'mridanga,' and 'nagaadaa.' There is only one term for all above-mentioned musical instruments in English which is 'drum.'

2.5 Solutions

These were the steps taken to find solutions for the above-mentioned problems:

Fig. 2.1 Hypernymy linkages for kinship relations



2.5.1 Kinship Relations

As mentioned above in challenges section that Hindi concept of *Bhatijaa* could not be directly linked to some Indian languages where the gender of the speaker and youngerness or elderness of the speaker or referent plays a major role. In this case, it was decided to link the synsets of the particular language to Hindi synset of *Bhatijaa* with a hypernymy relation, as the following diagram shows (Fig. 2.1).

2.5.2 Changes in Part-of-Speech

In Manipuri language, predicative adjectives cannot be linked as such, since in predicative position, adjectives become verb. Therefore, it has been decided that they will be changed to attributive adjectives and then the example sentences will be formed, e.g., *aada phajei nupii machaa eppii. vahaan sundar laDakii KhaDii hai.* 'A beautiful girl is standing there.'

There are Hindi concepts which cannot be linked with Sanskrit due to change in their part-of-speech in the language. However, since these words are available in instrumental cases of adjectives/nouns in the same sense, it has been decided to use the instrumental cases of sense instead of using them as adverb, e.g., 'shaantachetasaa,' 'shaantachitten,' 'shaantamanasaa,' and 'shaantahri' (instrumental cases of adjectives in Sanskrit which have been used as the adverb in language for 'shaantipoorvaka' of Hindi).

2.5.3 Different Terms in a Language for a Hindi Concept

In Kashmiri language, there are two different terms for water, one is for water in general and the other one is for drinking water, whereas Hindi has only one concept. The general term for water 'aaba' will be directly linked to the synsets for water 'paanii' of Hindi WordNet and 'tresh' which is a term for drinking water in Kashmiri will be linked to water 'paanii' of Hindi synset with a hypernymy

relation, e.g., general concept of water ‘aaba,’ drinking water ‘tresh.’ The first concept can be directly linked to water ‘paanii’ and the second one can be linked through hypernymy to water ‘paanii.’

2.5.4 Language-specific Synsets

To resolve the language-specific synsets, specific synset ids are allotted to each group. All groups will create synsets which are specific to their languages and culture in simple text file in the given id range. The groups will also generate parallel Hindi synsets for that concept in the same ids. This will be uploaded to the Hindi WordNet using the Hindi WordNet Data Entry Interface (description and snapshot in Sect. 2.6).

2.5.5 Challenges in Linking with English Synsets

To overcome the problem of Hindi–English synset linkage, this is the step taken. Hindi synsets can be linked with English synsets by two kinds of linkages, first is direct linkage and the other one is hypernymy linkage.

2.5.5.1 Direct Linkage

The Hindi synset having exact equivalents in English will be linked to them with direct linkage; e.g., ‘paanii’ is linked to water with direct linkage.

2.5.5.2 Hypernymy Linkage

Hindi synsets which cannot be linked directly to English concepts will be linked with hypernymy linkage; e.g., musical instruments which are part of Hindi WordNet, such as ‘tabalaa,’ ‘dhol,’ ‘mridanga,’ and ‘nagaadaa’ will be linked to ‘drum’ with a hypernymy relation (Fig. 2.2).

2.6 Tools for Hindi WordNet

Figure 2.3 shows the tools created at IIT Bombay for WordNet creation and maintenance along with their dependencies. Out of the tools shown in the diagram, the tools which help in WordNet creation in Hindi and other languages are discussed below.

Fig. 2.2 Hypernymy linkage in action

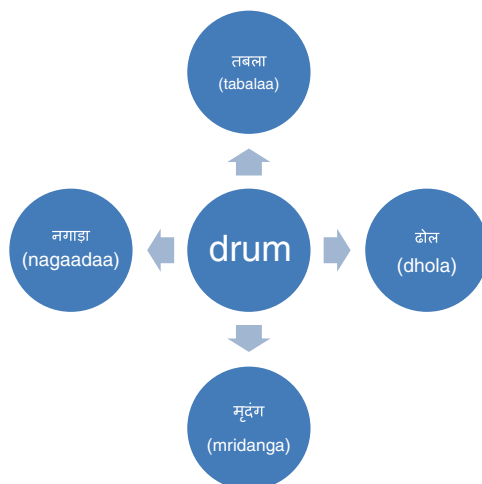
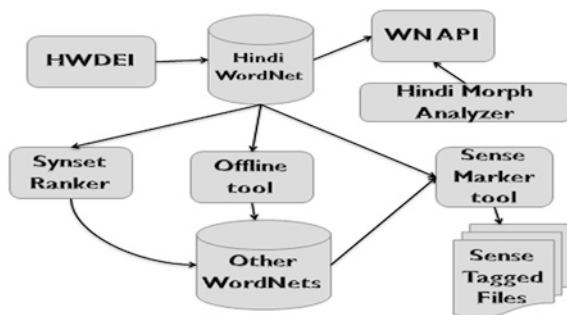


Fig. 2.3 All the tools and their dependencies



2.6.1 Hindi WordNet Data Entry Interface Tool

To facilitate a simple GUI-based synset insertion point for linguists working at IIT Bombay, this data entry interface was created. The interface allows the linguists to insert or modify the Hindi synsets easily.

This tool is designed for creating language-specific synsets and was originally created for Hindi language only. It is database-based tool, and the data entered using this tool are directly updated in Hindi WordNet database maintained at IIT Bombay. For this reason, this is used only within IIT Bombay by the linguistic team working on generation of Hindi WordNet (Fig. 2.4).

A similar interface was later on designed for Marathi WordNet as well, as the linguistic team for Marathi language is also part of IIT Bombay NLP group. This tool has facilities for faster lookup of category (nouns, adjectives, verbs, etc.)-specific search and includes options for specifying relations between synsets. The

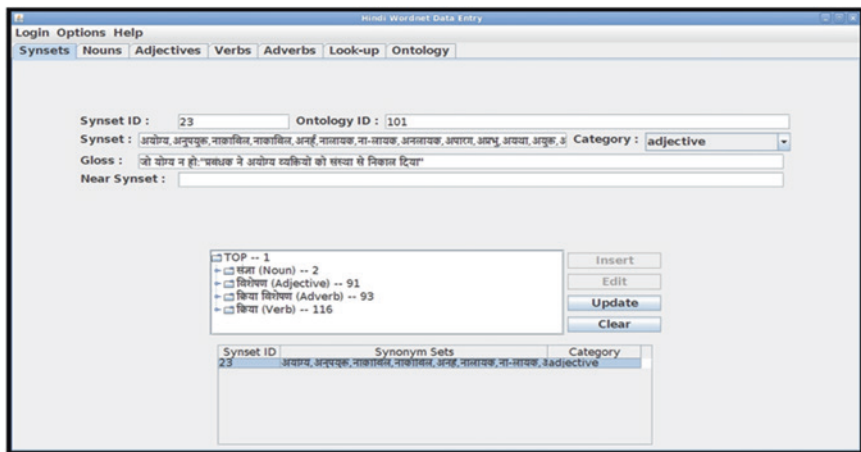


Fig. 2.4 Hindi WordNet data entry interface

tool allows the users to search for existing synsets using either the synset IDs or the synset words. It also keeps track of the ontology ID, category, and other fields which are related to the synset. It provides a facility of finding all synsets which contain a specific word or pattern. This is useful in cases where there are many synsets corresponding to a word or pattern, and the exact synset id is to be found out. The tool also includes user-friendly options for changing the font size, feel and look, etc. This tool is developed in java in order to make it as platform independent and works for Indic languages on platforms which support Unicode.

2.6.2 Off-line Tool

The off-line tool is a java-based open-source tool created for faster creation of Indic WordNets using Hindi WordNet as a pivot language. This tool provides a very similar feel to that of the online interface available for the Hindi WordNet and is publicly downloadable. The tool was created at IIT Bombay. Juxtaposed to HWDEI tool explained earlier, the off-line tool was created as a file-based tool for creation of WordNets. The tool uses Hindi WordNet as a pivot and provides a rapid way of constructing synsets for any language for the linguists (Fig. 2.5).

As shown in the screenshot, the left-side pane of the tool is where the Hindi synsets get loaded. The right-side pane is to be filled by the user in order to create synset in the target language (the screenshot is showing Sanskrit as the target language).

The tool provides some easy configuration options for setting the source and the target languages. By default the source language is Hindi (as it is the pivot language), but this can be changed to reflect any language of user's choice. The tool

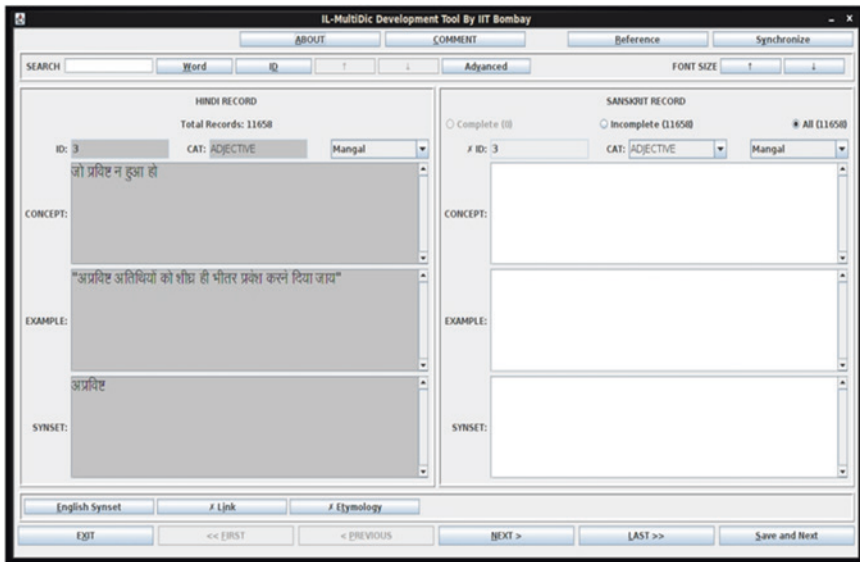


Fig. 2.5 Off-line tool

by default also allows the user to provide the English synset file (for reference) if it exists. Once the Concept, Example, Synset (words), Link, and Etymology fields are filled up for the target language, the synset is considered to be complete. This count helps the end user to gather data statistics and also to easily navigate through the incomplete synsets.

The latest version (v2.1) of the tool provides option for Secure Shell (SSH) Synchronization. This allows multiple users from same linguistic team to work on parallel for the same target language and then enables them to merge their work on a server through which the communication is done using SSH.

This tool provides standard options of changing the font size, navigation options, options for synset level comments, etc.

2.6.3 Synset Categorization Tool

As per the new distribution of synsets as decided in the 2nd IndoWordNet Workshop, there are six categories (ranks) of synsets for the Indian languages: Universal, Pan-Indian, Family-specific, Language-specific, Rare, and Synthesized. The Synset Categorizer or Ranker tool helps linguists categorize the synsets into one of the above six categories.

The Synset Categorizer tool is a very significant tool for IndoWordNet because synset making happens in a prioritized way in the following order: Universal → Pan-Indian → Family-specific → Language-specific → Rare → Synthesized (Fig. 2.6).

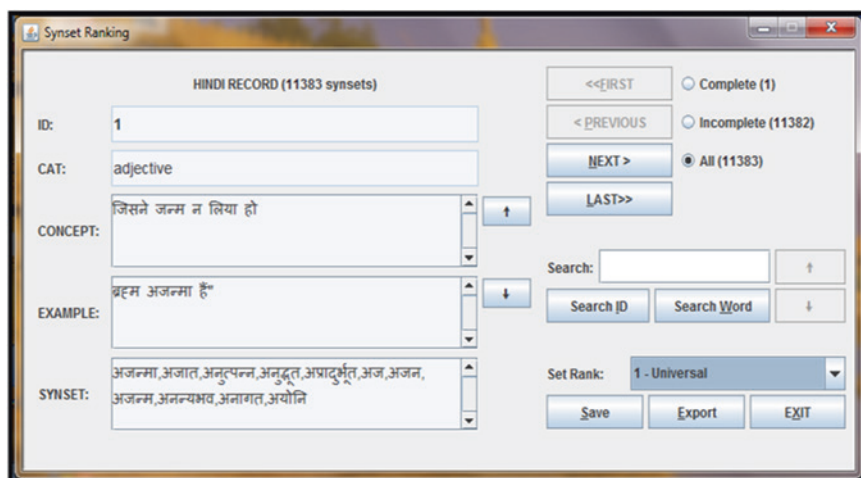


Fig. 2.6 Synset categorization tool

Besides, it automatically imposes a preliminary ontological structure on the synsets and if synsets are completed in this way, natural lexemes of the language get covered early, and the ambiguity that might have cropped up later is reduced at the initial stage.

The design for the tool is rather simple, compared to the previous tools. The tool shows the language synset in the left-side pane. The linguist's job is limited to deciding the category for this synset from the aforementioned categories.

Like off-line tool, this is a file-based tool and portable as it is written in Java. Since the tool follows the same syntax for the input file as that of the off-line tool, the file generated by off-line tool can be directly used for categorizing the synsets of that particular language.

As of now, for the task of identifying Universal and Pan-Indian synsets, the tool was modified to decide whether a particular Hindi synset is available in particular language, and the tool was distributed to all linguistic groups under IndoWordNet family to make a decision.

2.6.4 Morph Analyzer Tool

Morphology plays a crucial role in the working of various NLP applications. Whenever we run a spell checker, provide a query term to a Web search engine, explore translation or transliteration tools, use online dictionaries or thesauri, or try using text-to-speech or speech recognition applications, morphology works at the back of these applications.

Natural Language Processing (NLP) systems aim to analyze and generate natural language sentences and are concerned with computational systems and their interaction with human language. Morphology accounts for the morphological properties of

languages in a systematic manner, enabling us to understand how words are formed, what their constituents are, how they may be arranged to make larger units, what are the semantic and grammatical constraints involved, and how morphological processes interact with syntactic and phonological ones. An analysis of the inflectional morphology of Hindi has been presented here in the theoretical framework of Distributed Morphology, as discussed by Halle and Marantz (1993, 1994); Harley and Noyer (1999). The theory has been used to develop the rules required to analyze and describe the various inflectional forms of Hindi words. Our tool takes an inflected word as input and outputs its set of roots along with its various morphological features using the output of the stemmer. The suffixes extracted by the stemmer are used to get the various morphological features of the word: gender, number, person, case, tense, aspect, and modality. The tool consists of two parts—Stemmer, which takes inflected word as input and stems it, to separate root and suffix and Morphological Analyzer, which takes pair as input and outputs a set of features along with the set of roots (Bahuguna et al. 2014).

2.6.5 *Sense Marker Tool*

Annotation plays a key role in today's NLP scenario. And, one of the toughest annotation tasks is sense marking. In a given text, the occurrence of a particular word will correspond to only one sense and assigning the word with the correct sense from Hindi WordNet (or any other WN) is sense marking.

For machine translation of English to any Indian language, word sense ambiguity is a prominent issue. A huge amount of data needs to be sense-marked accurately by humans using an authentic and standard lexicon is essential (Fig. 2.7).

The sense marker tool is a Graphical User Interface-based tool which uses Java for manual sense marking. It can be used for any language, provided that the language has its own data base and morph analyzer. The tool displays the different senses of a word with gloss and entries of each synset to which the word belongs. It allows the user to select the correct sense of the word from among all the senses. Word can be tagged by just a single click on the correct synset (Fig. 2.8).

2.7 Web Interface

The current Hindi WordNet interface conforms to the Web standards and the latest UI/UX designs. The interface boasts of a 'card-ed' interface where each synset along with all its properties is shown in the form of a color-coded card. The cards are color coded according to their part-of-speech categories.

The interface also supports all the handheld devices and can be viewed at varied resolutions. In this era when a lot of Internet browsing is done through mobile devices, it is a preliminary need of Web-based design to be mobile supportive. Hence, we have modified the interface to support all such devices and, thus, all kinds of resolutions. The current interface not only does search for words based

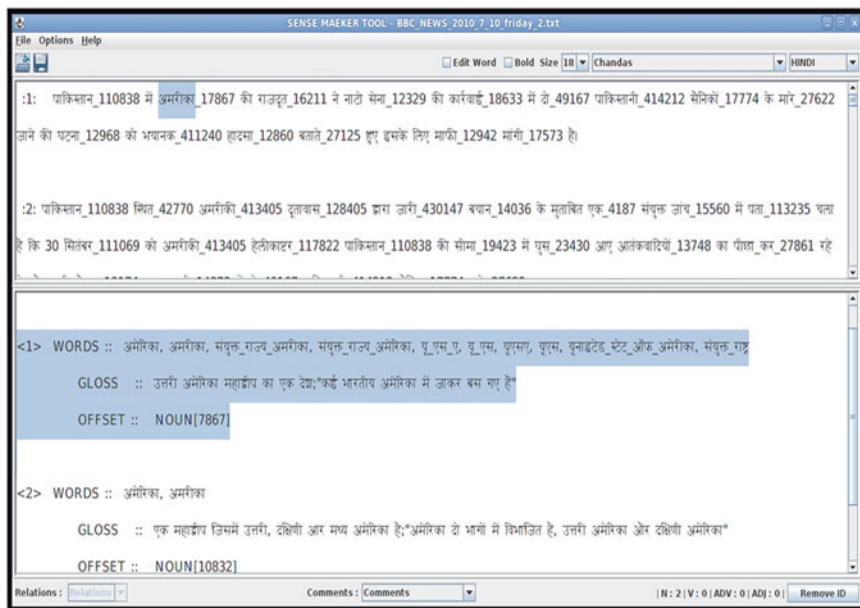


Fig. 2.7 Sense marker tool

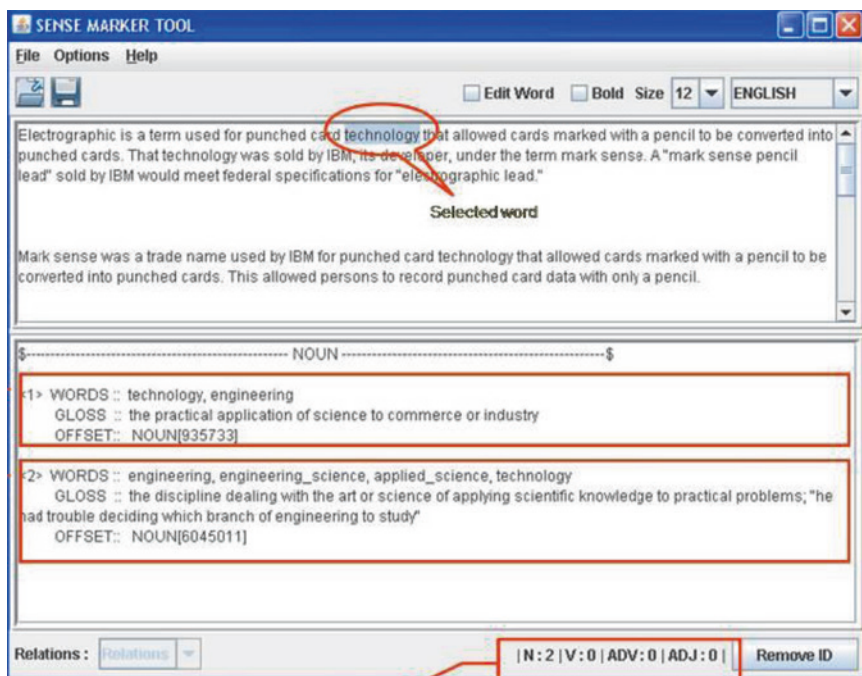


Fig. 2.8 Sense marker tool

www.cfilt.iitb.ac.in/~wordnet/hindinw/en.php

हिन्दी शब्दतंत्र (Hindi Wordnet)

हिंदी शब्द संकल्पनाघोष (A Lexical Database for Hindi)

Hindi Wordnet Introduction Wordnets Downloads References Feedback Login

✓ : Phonetic Transliteration

शुद्धता खोजें (Search)

देवनागरी कुञ्जीयरत (Devanagari Keyboard)

उदाहरण Examples सहायता Help **पनिष्क्रिया दें Give feedback** पनिष्क्रियाओं देखें Previous feedbacks पुराना इंटरफेस Previous interface मराठी शब्दघण्टा Marathi Wordnet संस्कृत शब्दघण्टा Sanskrit Wordnet हिन्दी-अंग्रेजी शब्दघोष Hindi-English Dictionary

इंडो-अंग्रेजी शब्दघण्टा Indo Wordnet Online इंडो-अंग्रेजी विजुअलाइजर Indo Wordnet Visualizer सी.एफ.आई.एल.टी. मुखपृष्ठ CFILT Home हिन्दी शब्दतंत्र मुखपृष्ठ Hindi Wordnet Home

Total Unique Words: 100273 | Total Synsets: 39992 | Total Linked Synsets: 25853 | Bilingual Mappings: 2447 | Last Updated: 14 Jul 2015

[illegible]

Fig. 2.10 HWN web interface

in/out, expand, and fix nodes for better visibility. The screenshot for the Hindi word ‘*diiwaara*’ is shown in the figure given below. This interface is very useful in various NLP applications, viz., WordNet Validation, Semantic Relatedness, Word Sense Disambiguation, Information Retrieval, and Textual Entailment (Fig. 2.12).

2.9 Hindi WordNet Mobile Application

In this era of handheld mobile devices, there is a great need to make available the traditional Web services as mobile applications (Kanojia et al. 2016) which are extremely popular. Web browser-based interfaces are available but are not suited for mobile devices, which deters people from effectively using WordNets. We developed mobile applications for the Android and iOS platforms for Hindi WordNet which allows users to search for words and obtain more information about them along with their translations in English and other Indian languages.

India is a country in the world with massive language diversity. According to a recent census in 2001, there are 1365 rationalized mother tongues, 234 identifiable mother tongues and 122 major languages. Of these, 29 languages have more than a million native speakers, 60 have more than 100,000 and 122 have more than 10,000 native speakers. With this in mind, the construction of the Indian WordNets, the IndoWordNet (Bhattacharyya 2010) project was initiated which was an effort undertaken by over 12 educational and research institutes headed by IIT Bombay.

It is common knowledge that Web sites such as Facebook, Twitter, LinkedIn, and Gmail can be accessed using their Web browser-based interfaces but the mobile applications developed for them are much more popular. This is a clear indicator that browser-based interfaces are inconvenient which was the main motivation behind our work. We studied the existing interfaces and the WordNet databases and developed applications for Android, iOS, and Windows Phone platforms. We believe that such an application can be quite helpful in a classroom educational scenario, where students, often belonging to different cultural and linguistic background, would be able to access this application as a dictionary for multiple Indian languages.

The current application for Hindi WordNet has been launched on Android and Windows Phone stores, which are available for free download. It can be used to search for words using Hindi as the pivot language and then results are multilingual connected to each other based on the same sense. Another possible use of this application can be where tourists traveling to India use this application for basic survival communication needs.

Hindi is fairly new to the Web, and despite of standard UTF encoding of characters, there remain a few steps to be taken to sanitize the input for WordNet search. We inculcated steps such as nukta normalization and morphological analysis of the input word to ease the search in our mobile application. The results returned by the server are interpreted by the application pages and displayed in a very simplistic manner. We display all synsets for each part-of-speech and all senses of that word and initially showing the synset words, gloss, and example.

These senses are categorized by their part-of-speech categories. We have conformed to the principles of good user interface design and provided for an incremental information display.

We plan to make efforts toward improving this application to enable searching for words belonging to all languages which have a common interface via language detection. We also plan to inculcate Word Suggestions as they are being typed so that the user is presented with better lexical choices (Fig. 2.13).

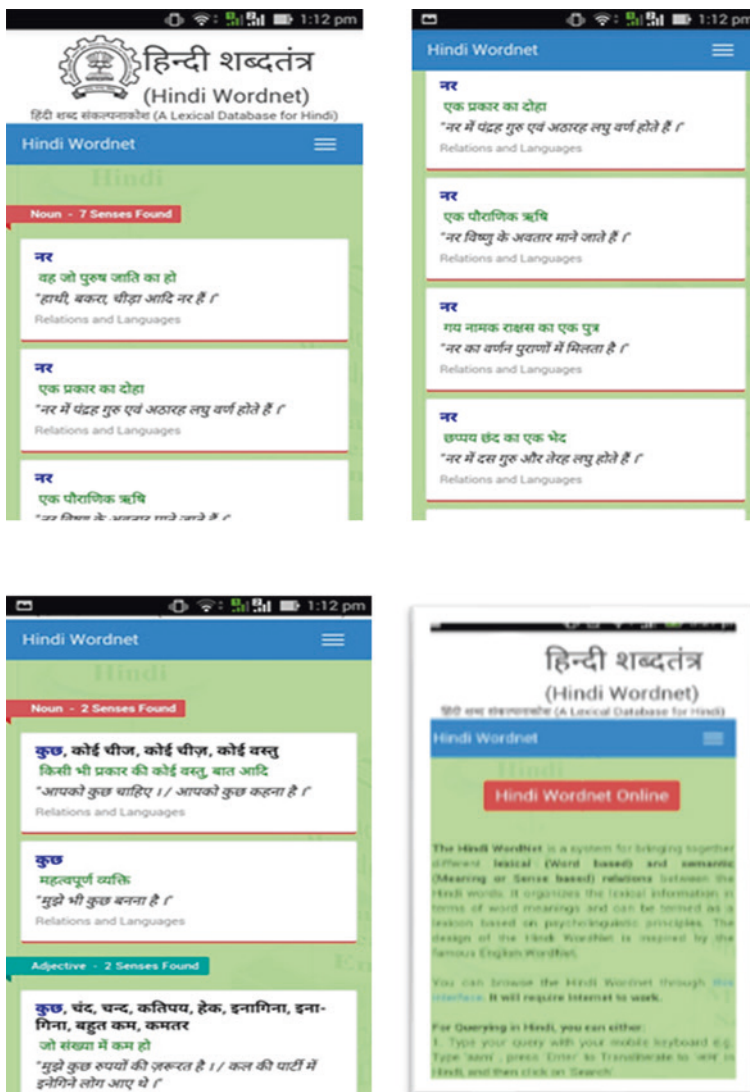


Fig. 2.13 Hindi WordNet mobile application

Fig. 2.14 Online synset creation interface

2.10 Online Synset Creation Interface—Synskarta

Synskarta (Redkar et al. 2014) is an online interface designed for creating synsets from source language to target language by following the expansion approach. It is a centralized Web-based system which uses relational database to store and maintain synset and related data. The difficulty of maintaining data in flat files is taken care of by Synskarta. It is developed using PHP and MySQL. The IndoWordNet database structure is used for storing and maintaining the synset data while IndoWordNet APIs are used for accessing and manipulating these data (Fig. 2.14).

2.10.1 Some of the Salient Features of Synskarta are as Follows

- **User Registration Module**—This module allows the system administrator to create user profiles and provide necessary access privileges to his/her role. The user can login using the access privileges provided to him and accordingly the user interface is displayed to that particular user.
- **Configuration Module**—This module sets all the necessary parameters such as source language and target language and enables or disables certain features such as Source, Domain, Linking, Comment, and References.
- **Main Module**—This module allows the user to enter data in the target language panel by referring to data in the source language panel. The source panel and the target panel vertically divide the main module into two equal sized panels. Following are the major components of this module:
- **Source language panel**—This panel is placed on the left of the screen which has fields for synset id, POS category, gloss, example(s), and synonyms of the source language synset.

- Target language panel—This panel is placed on the right of the screen. This panel has non-editable fields such as synset id, POS category and editable fields such as gloss, example(s), and synonym(s) of the target language synset. The user is expected to enter the data in these editable fields.
- Search—User can search a synset either by entering 'synset id' or a 'word' in a synset.
- Advanced Search—Here user is allowed to search synset data by entering various parameters, such as POS category and words appearing in gloss or in example.
- Comment—User can comment on a particular synset being translated.
- English Synset—User can check the corresponding English synset for better clarity in translation process.
- Navigation Panel—This panel allows the user to navigate between synsets. Button 'Save & Next' allows inserting or updating a current target synset and the data are directly stored in the IndoWordNet database.
- Vindication—This feature allows the user to record the special feature of a particular word in a current synset.
- Source—This feature allows the user to record the information about source of the synset.
- Domain—This feature allows the user to record the information about domain of the synset.
- Linking—Many-to-many linking of words is supported in this feature.
- Quotations—The feature is to add quotations as additional examples are supported.
- Root Verb—This feature allows the user to enter the root verb of a given word.
- Feedback—Feedback related to the tool and its features are captured here.
- Word Options—Provides specific features which can be specific to a particular word in the synset. There are features which are specific to the Sanskrit language. Some of these features can also be applicable to other languages. As far as Sanskrit is concerned, some of the features are Indication of Word Type, Indication of Accent, Identification of Gender, Indication of Proverbs, Indication of Class, Etymology, Indication of Transitivity, Indication of Ittva, Indication of Pada, and Indication of Verbal Root Types.

2.10.2 Advantages and Limitations of Synskarta

2.10.2.1 Major Advantages of Synskarta are as Follows

- Centralized system—Online access from anywhere in the world
- No data redundancy and inconsistency.
- No text files to maintain data
- Faster processing and updating
- Multiple users can work at the same time
- Can be used by all the language WordNets

Table 2.1 Synset linkage status

No.	Languages	Linked status
01	Assamese	14,958
02	Bodo	15,785
03	Manipuri	16,351
04	Nepali	11,713
05	Tamil	25,431
06	Telugu	21,091
07	Malayalam	27,180
08	Kannada	20,033
09	Kashmiri	29,469
10	Konkani	32,370
11	Marathi	30,817
12	Sanskrit	23,176
13	Gujarati	35,599
14	Panjabi	32,364
15	Bengali	36,346
16	Urdu	34,280
17	Oriya	35,284

2.10.2.2 Disadvantage

The major limitation of Synskarta is it is heavily dependent on Internet access or networking.

2.11 Linkage Statistics

The table shows linkage status of synsets of other language WordNets with Hindi WordNet as mentioned earlier, and these seventeen (17) languages follow Hindi WordNet as pivot for creation of WordNets in the respective language (Table 2.1).

The status mentioned in the table shows the counts of synsets from these languages which are linked to Hindi language as of July 2015.

2.12 Conclusion

In this paper, we have discussed problems faced in categorization of Hindi synsets and the different categories chosen. The paper presents challenges faced while categorizing and in linking Hindi synsets with different languages. The solution suggested accommodates culture-specific concepts such as different types of cuisines of different places, folk dances, and folk songs in a particular language WordNet.

A suggestion of using Hypernymy linkage has been proposed to solve the problem of linkage of culture specific or uncommon synsets in different languages. Current linkage statistics for all languages in IndoWordNet has been mentioned. Near-future plan is to link language-specific synsets with Hindi WordNet.

The paper also presented different tools which were developed for WordNet synset data entry, linkage, categorization, etc.

Acknowledgments The support of the Dept. of Information Technology, Govt. of India, toward the WordNet development effort through *Indradhanush* project is thankfully acknowledged.

Appendix: Some of the Universal Synsets Selected by IndoWordNet Members

Id	Synset	Id	Synset	Id	Synset	Id	Synset
27	<i>moorkha</i> (fool)	273	<i>masooDaa</i> (gum)	561	<i>tarjanii</i> (index finger)	779	<i>jyeshTha</i> (name of a lunar month)
30	<i>yogya</i> (capable)	278	<i>hiiraa</i> (diamond)	562	<i>anaanikaa</i> (ring finger)	780	<i>saavana</i> (name of a lunar month)
31	<i>sabhya</i> (decent)	293	<i>Boonda</i> (drop)	568	<i>arahara</i> (toor pulse)	781	<i>bhaadrapada</i> (name of a lunar month)
44	<i>apamaan</i> (insult)	298	<i>khoona</i> (blood)	591	<i>maaranaa</i> (to hit)	782	<i>pousha</i> (name of a lunar month)
46	<i>sammaana</i> (respect)	332	<i>pinjaraa</i> (cage)	592	<i>latiyaanaa</i> (kick)	787	<i>chaadara</i> (bed sheet)
47	<i>iishwar</i> (God)	335	<i>choohaa</i> (mouse)	600	<i>jaala</i> (net)	788	<i>asatya</i> (lie)
51	<i>achaanaka</i> (suddenly)	344	<i>sariiyaa</i> (rod)	604	<i>mastaka</i> (fore head)	792	<i>dhanusha</i> (bow)
75	<i>sadguNa</i> (virtue)	345	<i>kalama</i> (pen)	605	<i>cheharaa</i> (face)	798	<i>dvaara</i> (door)
120	<i>prema</i> (love)	346	<i>kaagaja</i> (paper)	610	<i>tanaa</i> (stem)	801	<i>paradaa</i> (curtain)
121	<i>sneha</i> (love)	370	<i>baansurii</i> (flute)	617	<i>kala</i> (yesterday)	802	<i>bichhounaa</i>
129	<i>anuvaad</i> (translation)	373	<i>paasa men</i> (nearby)	623	<i>laala ranga</i> (red)	806	<i>kambala</i> (blanket)
142	<i>ghriNaa</i> (hatred)	406	<i>sitaara</i> (name of a string instrument)	624	<i>haraa</i> (green)	808	<i>maagha</i> (name of a lunar month)

155	<i>oura</i> (other)	409	<i>taanapoora</i> (name of a string instrument)	625	<i>niilaa</i> (blue)	822	<i>puraskaara</i> (prize)
171	<i>aparadha</i> (crime)	417	<i>motaa</i> (plump)	630	<i>poornimaa</i> (full moon)	849	<i>sinha</i> (lion)
172	<i>aparaadhii</i> (criminal)	422	<i>paira</i> (leg)	631	<i>madhya raatri</i> (midnight)	858	<i>aadamii</i> (man)
203	<i>baraamadaa</i> (veranda)	444	<i>biskuta</i> (biscuit)	632	<i>kala</i> (tomorrow)	890	<i>taaraa</i> (star)
208	<i>graha</i> (home)	452	<i>mandira</i> (temple)	635	<i>makkhii</i> (fly)	893	<i>khulaa</i> (uncovered)
217	<i>lohaa</i> (iron)	464	<i>moorti</i> (statue)	642	<i>tahanii</i> (twig)	965	<i>qaanoona</i> (law)
225	<i>budha</i> (Mercury)	470	<i>rangamancha</i> (theater_ stage)	643	<i>konpala</i> (foliage)	976	<i>taalaa</i> (lock)
226	<i>shukra</i> (Venus)	473	<i>dhaagaa</i> (thread)	644	<i>shaakhaa</i> (tree branch)	984	<i>darshaka</i> (spectator)
227	<i>brihaspati</i> (Jupiter)	474	<i>dhaatu</i> (metal)	647	<i>haddii</i> (bone)	985	<i>naaka</i> (nose)
228	<i>shani</i> (Saturn)	476	<i>duma</i> (tail)	648	<i>chaitra</i> (name of a lunar month)	987	<i>kaana</i> (ear)
229	<i>varuNa</i> (Neptune)	491	<i>bhujaa</i> (arm)	652	<i>avastha</i> (state)	990	<i>nathunaa</i> (ala)
231	<i>pradesha</i> (province)	492	<i>poshaaka</i> (clothing)	661	<i>pasalii</i> (rib)	991	<i>niraashaa</i> (hopelessness)
233	<i>zillaa</i> (district)	504	<i>kalaaii</i> (wrist)	667	<i>polaa</i> (hollow)	1011	<i>naraka</i> (hell)
235	<i>ronaa</i> (cry)	505	<i>kuhanii</i> (elbow)	679	<i>haara</i> (defeat)	1012	<i>damaa</i> (asthma)
236	<i>gaanna</i> (sing)	511	<i>ulkaa</i> (meteoroid)	720	<i>mitrataa</i> (friendship)	1013	<i>prasannataa</i> (cheerfulness)
247	<i>satyavaadii</i> (honest)	521	<i>jhandaa</i> (flag)	749	<i>koyala</i> (cuckoo)	1017	<i>Indradhanusha</i> (rainbow)
268	<i>riiDa</i> (spine)	526	<i>vaakaii</i> (actually)	752	<i>pankha</i> (wing)	1029	<i>fena</i> (foam)
269	<i>shakti</i> (strength)	528	<i>spashta</i> (clear)	753	<i>choncha</i> (beak)	1036	<i>mahaavata</i> (mahout)
270	<i>munha</i> (mouth)	531	<i>praaNa</i> (spirit)	758	<i>phana</i> (hood)	1038	<i>Imalii</i> (tamarind)

271	<i>daanta</i> (tooth)	532	<i>naayikaa</i> (heroine)	762	<i>magar</i> (crocodile)	1045	<i>shaanti</i> (peace)
272	<i>taaloo</i> (palate)	558	<i>choolhaa</i> (stove)	778	<i>vaishaakha</i> (name of a lunar month)	1046

Note Hindi words are written in *italic* and English meanings are given in the brackets in the above table

References

- Bahuguna, A., Talukdar, T., Bhattacharyya, P., & Singh, S. (2014). HinMA: Distributed morphology based hindi morphological analyzer. In *International Conference on Natural Language Processing 2014 (ICON 2014)*, Goa University, Goa, India, 19–20 December 2014.
- Bhattacharyya, P. (2010). IndoWordNet. In *Lexical Resources Engineering Conference 2010 (LREC 2010)*, Malta, May, 2010.
- Chaplot, D. S., Bhingardive, S., & Bhattacharyya, P. (2014). IndoWordnet visualizer: A graphical user interface for browsing and exploring wordnets of Indian languages. In *Global WordNet Conference 2014 (GWC 2014)*, Tartu, Estonia, 25–29 January 2014.
- Halle, M., Marantz, A. (1993). Distributed Morphology and the Pieces of Inflection. In Kenneth Hale & S. Jay Keyser (Eds.), *The view From Building 20* (pp. 111–176). Cambridge: MIT Press.
- Halle, M., Marantz, A. (1994). Some key features of Distributed Morphology. In Andrew Carnie and Heidi Harley (Eds.), *MITWPL 21: Papers on phonology and morphology* (pp. 275–288). Cambridge: MITWPL.
- Harley, H., Noyer, R. (1999) Distributed morphology. *Glott international*, 4(4), 3–9.
- Kanojia, D., Dabre, R., & Bhattacharyya, P. (2016). Sophisticated Lexical Databases-Simplified Usage: Mobile Applications and Browser Plugins For Wordnets. In *Global WordNet Conference (GWC 2016)*, Bucharest, Romanian, 27–30 January 2016.
- Naravane, V. D. (1961). *Bharatiya Vyavahara Kosha: Solah Bhasao ka kosha*. Triveni Samgama. (in Hindi).
- Redkar, H., Paranjape, J., Joshi, N., Kulkarni, I., Kulkarni, M., & Bhattacharyya, P. (2014). Introduction to Synskarta: An online interface for synset creation with special reference to Sanskrit. In *International Conference on Natural Language Processing 2014 (ICON 2014)*, Goa, India, 19–20 December 2014.
- Saraswati, J., Shukla, R., Goyal, R. P., & Bhattacharyya, P. (2010). Hindi to English WordNet linkage: Challenges and solutions. In *Proceedings of 3rd IndoWordNet Workshop, International Conference on Natural Language Processing 2010 (ICON 2010)*, Indian Institute of Kharagpur, India, 8–11 December 2010.

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