Creation of Punjabi WordNet and Punjabi Hindi Bilingual Dictionary

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Computer Science

Submitted By:
Rekha Rattan
(800932016)

Under the supervision of:
Parteek Bhatia
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004
June 2011

Certificate

I hereby certify that the work which is being presented in the thesis entitled, "*Creation of Punjabi WordNet and Punjabi Hindi bilingual dictionary*", in partial fulfillment of the requirements for the award of degree of *Master of Engineering in Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of Mr. Parteek Bhatia and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Rekha

(Rekha Rattan)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Parteek Bhatia

(Mr. Parteek Bhatia)

Assistant Professor
Computer Science and Engineering Department,
Thapar University, Patiala

Countersigned by

(Dr. Maninder Singh)
Head 24/12/21
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. K. Mohapatra)
Dean (Academic Affairs)
Thapar University
Patiala

i

Acknowledgement

First of all, I would like to express my gratitude towards Thapar University**,** for providing me a platform to do my thesis work at such an esteemed institute.

I wish to express my deep gratitude to Mr. Parteek Bhatia, Assistant Professor, Computer Science and Engineering Department, Thapar University, Patiala for his valuable advice and guidance in carrying out my thesis.

I would like to thank Dr. Maninder Singh, Head, Computer Science and Engineering Department, Thapar University, Patiala who has been a constant source of inspiration for me throughout this work.

I am also thankful to all the staff members of the Department for their full cooperation and help.

I am Thankful to my family and all my friends for their blessings and moral support. Thanks for boosting me with their constant encouragement, support and confidence.

Last but not the least, I am thankful to God for providing me with the strength and ability to complete my work.

Rekha Rattan

Abstract

India is a multilingual country wherein people speak many different languages .Therefore a lot of work can be done in the fields of machine translation and cross lingual information processing. This acted as a motivation behind the building of IndoWordNet. IndoWordNet is a linked structure of WordNets of major Indian languages. Many Indian languages are using the expansion approach to develop their WordNets from the Hindi WordNet. Punjabi is the language of hundreds of millions of people in India, and is the religious language of all Punjabis around the world. Surprisingly little work has been done in the field of computerized language and lexical resources for this language. It is therefore worthy to build up a Punjabi lexical resource (WordNet) that can discover the richness of language Punjabi.

This thesis consists of a brief overview of WordNet and its underlying main beliefs in different languages. This thesis describes our approach towards building a lexical resource in Punjabi language. Punjabi WordNet is being developed from Hindi WordNet using the expansion approach. An algorithm for creating Punjabi Hindi bilingual dictionary by using Punjabi WordNet has also been proposed in this thesis work. A methodology for development of a web application for Punjabi Hindi bilingual dictionary has also been discussed in this thesis.

Table of Contents

Certificate **i**

Acknowledgement **ii**

Abstract **iii**

Table of Contents **iv-v**

List of Figures **vi-vii**

List of Tables **viii**

Chapter 1 Introduction **1-8**

1.1 Natural Language Processing 1

1.2 Challenges in Natural Language Processing 2

1.3 Role of WordNet in Natural Language Processing 3

1.4 IndoWordNet creation process 5

1.5 Differences between IndoWordNet and EuroWordNet 7

Chapter 2 Literature Review **9-24**

2.1 WordNet principle 9

2.1.1 Lexical matrix 9

2.1.2 Synset making 10

2.2 General methodology for WordNet creation 11

2.2.1 Comparison of merge and expansion approach for WordNet building 12

2.3 Creation of Punjabi WordNet from Hindi WordNet using expansion approach 12

2.4 Constituent elements of WordNet 14

2.4.1 Synset and concept 14

2.4.2 Relations in WordNet 15

2.4.2.1 Semantic relations 16

2.4.2.2 Lexical relations 19

2.5 Demonstration of all relations in WordNet 19

2.6 Existing electronic dictionaries in Punjabi 20

Chapter 3 Problem Statement **25-27**
3.1 Motivation 25
3.2 Gap analysis 26
3.3 Problem statement 26
3.4 Objectives 27
3.5 Methodology 27
Chapter 4 Design and implementation of Punjabi WordNet and Punjabi Hindi bilingual dictionary **28-39**
4.1 Creation of synsets for Punjabi using IL-Multidic development tool 28
4.1.1 Configuration of tool 28
4.1.2 Working of tool 29
4.2 Creation of Punjabi WordNet 33
4.3 Creation of Hindi to Punjabi dictionary 36
4.4 Creation of Punjabi to Hindi dictionary 39
Chapter 5 Experimental results **40-47**
5.1 Web interface for Punjabi WordNet 40
5.2 Web interface for Hindi to Punjabi dictionary 42
5.3 Web interface for Punjabi to Hindi dictionary 44
Chapter 6 Conclusions and Future Scope **48-49**
6.1 Conclusions 48
6.2 Future scope 48
References **50-51**
List of Publications **52**

List of Figures

Figure 1.1	Linked IndoWordNet structure	7
Figure 2.1	Punjabi WordNet synset	13
Figure 2.2	Hypernymy/Hyponymy relations	17
Figure 2.3	Meronymy/Holonymy relations	18
Figure 2.4	Relations for synset of ਘਰ (Home)	20
Figure 2.5	Web interface for Punjabi Kosh	21
Figure 2.6	Web interface for Punjabi English dictionary	22
Figure 2.7	Web interface showing different senses for word "*life*" in Punjabi	23
Figure 2.8	Web interface showing different senses for word "*health*" in Punjabi	23
Figure 2.9	Punjabi Shabdkosh	24
Figure 4.1	Configuring Il-Multidic development tool	29
Figure 4.2	Tool for creating standardized lexical data	30
Figure 4.3	Source files of Hindi synstes	31
Figure 4.4	Output file containing entries for Punjabi words	33
Figure 4.5	Final result list for the word "ਵਸਤੂ"	34
Figure 4.6	Flowchart for Punjabi WordNet	35
Figure 4.7	Final result list for the word "आम"	37
Figure 4.8	Flowchart for Hindi to Punjabi dictionary	38
Figure 5.1	Interface showing Punjabi WordNet showing Punjabi keypad	40
Figure 5.2	Interface showing first sense of the word "ਪ੍ਰਤੀਕੂਲਤਾ"	41
Figure 5.3	Interface showing second sense of the word "ਪ੍ਰਤੀਕੂਲਤਾ"	41
Figure 5.4	Interface showing Hindi keypad for Hindi-Punjabi dictionary	42
Figure 5.5	Interface showing first sense of the word "आम" in Punjabi	42
Figure 5.6	Interface showing second sense of the word "आम" in Punjabi	43
Figure 5.7	Interface showing third sense of the word "आम" in Punjabi	43

Figure 5.8	Interface showing fourth sense of the word “ਆਸ” in Punjabi	44
Figure 5.9	Web interface for Punjabi to Hindi dictionary	45
Figure 5.10	Interface showing first sense of the word “ਗਿਆਨ” in Hindi	45
Figure 5.11	Interface showing second sense of the word “ਗਿਆਨ” in Hindi	46
Figure 5.12	Interface showing third sense of the word “ਗਿਆਨ” in Hindi	46
Figure 5.13	Interface showing fourth sense of the word “ਗਿਆਨ” in Hindi	47
Figure 5.14	Interface showing fifth sense of the word “ਗਿਆਨ” in Hindi	47

List of Tables

Table 1.1	WordNets of different languages and institutes developing them	6
Table 2.1	Illustrating the concept of lexical matrix	10
Table 2.2	Example lexlcal matrix	10
Table 2.3	Semantic relations in WordNet	16
Table 2.4	Examples of synonyms	16

Chapter 1
Introduction

The information age has been characterized by the development and convergence of computing, telecommunications and multilingual information systems. This has resulted in the availability of large volumes of information in electronic media which is in the data presentation formats typical of computer systems, but whose natural language form, is more suited for human users than computer systems. This has in turn encouraged the development of technologies that would solve this problem and can help in accessing this information more efficiently and quickly. Natural Language Processing (NLP) provides tools and techniques that can allow the implementation of natural language-based interfaces to computer systems that can enable communication between man and machine in natural languages [1].

These techniques also enable people to organize, extract and use the knowledge contained in these huge collections of natural language electronic data. Examples of Language Technology (LT) applications include Machine Translation (MT), Information Extraction (IE), Information Retrieval (IR), document classification and summarization, speech recognition and synthesis *etc.* Lexical resources have become important basic tools within NLP and related fields. WordNet is a very rich source of lexical knowledge. A WordNet is a lexical database in which nouns; verbs, adjectives and adverbs are organized in a conceptual hierarchy, linking semantically and lexically related concepts.

1.1 Natural Language Processing

Natural language processing (NLP) is a subfield of artificial intelligence and linguistics. It studies the problems of automated generation and understanding of natural human languages. Natural language generation systems convert information from computer databases into human language, and natural language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate [1]. NLP has significantly overlapped with the field of computational linguistics, and is often considered a sub-field of artificial intelligence.

1.2 Challenges in Natural Language Processing

Natural Language Processing includes the tasks which deal with human languages. There are various challenges in Natural Language Processing that are to be dealt with. The following are the problems faced while processing human languages:

- **Text segmentation:** Some written languages like Chinese, Japanese and Thai do not have signal word boundaries either, so any significant text parsing usually requires the identification of word boundaries, which is often a difficult task.

- **Speech segmentation:** In most spoken languages, the sounds which represent successive letters merge into each other; therefore it becomes a difficult task to convert an analog signal to distinct characters. Moreover, in natural speech there are hardly any pauses between successive words, hence if a system has to locate these boundaries then it must be able to consider grammatical and semantic constraints, and also the context.

- **Word sense disambiguation**: Words can have various senses in which it can be used, *i.e.*, words can have more than one meaning, we have to select the meaning which makes the most appropriate sense in a particular context in which it is being used. A machine is able to distinguish between two senses of a word with the help of a "universal encyclopaedia"; WordNet serves the purpose of this universal encyclopaedia. Word sense disambiguation has many commercial applications, among which are the intelligent dictionaries, thesauri and grammar checkers. For example, students looking for definitions or synonyms of unfamiliar words are often confused by the definitions/synonyms for contextually inappropriate senses. Once the correct sense has been identified for the currently highlighted word in the context, an intelligent dictionary/thesaurus would list only the definition(s) and synonyms(s) appropriate for the actual sense [2].

- **Syntactic ambiguity**: The grammar for natural languages is ambiguous, *i.e.*, most probably there will be multiple parse trees possible for any given sentence. Thus, semantic and contextual information about a sentence is required in order to choose the most appropriate sense. Specific problem components of syntactic

ambiguity include sentence boundary disambiguation. For the structural ambiguity, consider the sentence: “The man saw the girl with a red hat”. This sentence is ambiguous as it can be interpreted in two different ways by the machine: The man saw the girl who was wearing a red hat or, the man saw the girl with the help of the red hat. However, the sentence “The man saw the girl with a red hat” is not ambiguous for any human reader, one can understand that a hat cannot be used to see, while it is difficult for a computer to interpret the same meaning correctly.

1.3 Role of WordNet in Natural Language Processing

Natural language processing is essential for dealing efficiently with the large quantities of text available online. For example, NLP is used in information retrieval, in text processing and in machine translation. Another essential function is helping the user with query formulation through synonym relationships between words and hierarchical and other relationships between concepts [1]. WordNet supports both of these functions.
Lexical resources have become fundamental tools within NLP and its related fields. The range of resources available to the researcher is diverse and vast - from simple word lists to complex dictionaries and thesauruses. The resources contain a whole range of different types of explicit linguistic information presented in different formats and at various levels of granularity.

Applications of WordNet in NLP

WordNet is used as a rich source of lexical information in many applications which are discussed below.

- **Human Language Technology and Artificial Intelligence**: WordNet has become a very useful resource in the human language technology and artificial intelligence. WordNet provides the general lexico-semantic information which helps in open-domain text processing.

- **Trans-lingual applications**: The development of WordNets in several other languages extends this capability to trans-lingual applications, enabling text

mining across languages. For example, in Europe, English WordNet has provided the starting point for the development of a multilingual database for several European languages which is called the EuroWordNet project.

- **Word Sense Disambiguation**: Word Sense Disambiguation is regarded as one of the most interesting and longest-standing problems in Natural Language Processing [3]. It is the process of determining which sense of a word is the intended sense in a particular context. For example, in the sentence, "John took his wife to the annual ball", a human could easily understand that '*ball'* is being referred as dance form. But it becomes difficult for software to detect which sense of '*ball'* was intended. Word sense disambiguation involves selecting the intended sense of a word for a predefined set of words, with the help of a machine-readable dictionary, such as WordNet. Word sense disambiguation which means a task of removing the ambiguity of word in a specific context is important for many NLP applications such as: Machine Translation, Speech processing, Text processing and Grammatical analysis [4].

- **Assessment of semantic similarity**: It has proved to be essential for a variety of Natural Language Processing (NLP) tasks, including syntactic disambiguation, word sense disambiguation, selection of a suitable translation equivalent, query expansion and document indexing in Information Retrieval (IR). The semantic similarity between two words is calculated on the basis of taxonomical associations. If we are given two word senses W1 and W2, their similarity is calculated as a function of their belongingness to more general semantic classes [7]. This approach again requires basic lexico-semantic information which is provided by WordNet.

- **To quantify the relatedness of two words**: Sometimes we wish to measure the relatedness of two word senses, where a word sense is a specific meaning of a particular word. For example, the word '*ball*' has several senses, it could mean an object used in games, a famous dance form, or a pitch in baseball. A specific method for quantifying how similar two word senses are is known as a measure of semantic relatedness [7].

Due to the significant increase in the use of lexical databases, WordNet becomes one of lexical databases that are widely used as a lexical information source for many applications: *e.g.* information retrieval, text classification, semantic disambiguation *etc* [5]. Today WordNet is available in many different languages and platforms, having different features and interfaces depending on the various objectives for which they were built. WordNet is presented as a software package, which can bound together the data (files in some codification) and the applications. Recent projects like EuroWordNet, BalkaNet and MultiWordNet started the development of WordNets for many other languages, thus leading to multilingual processing. Further, WordNets are now being developed world-wide; the Global WordNet Association maintains a list of existing WordNets which currently contains more than 30 languages. The design of English WordNet has been extended to other languages such as Dutch, Italian, Spanish, Hungarian and Chinese.

Hindi WordNet was the first WordNet developed for any of the Indian languages and now WordNets are being developed for major Indian languages under the IndoWordNet project from Hindi WordNet by following the expansion approach. For Punjabi language, the Punjabi WordNet is not available till date. Construction of WordNet lexical database, Punjabi WordNet, is a long term project. Punjabi WordNet is being developed at Thapar University under the IndoWordNet project.

1.4 IndoWordNet creation process

Seeing the enormous potential of WordNet, 16 out of 22 official languages of India, have started making their WordNets under the leadership of IIT Bombay. These languages are: (1) Hindi, (2) Marathi, (3) Konkani, (4) Sanskrit, (5) Nepali, (6) Kashimiri, (7), Assamese, (8) Tamil, (9) Malyalam, (10) Telugu, (11) Kannad , (12) Manipuri and (13) Bodo, (14) Bangla19, (15) Punjabi and (16) Gujarati [8]. These languages cover the length and breadth of India and are spoken by about 900 million people. Table 1.1 shows the WordNets and the corresponding institutes developing them.

Table 1.1: WordNets of different languages and institutes developing them [8]

WordNet	Language Institute(s)
Assamese	Guhati University, Assam
Bengali	Indian Statistical Institute,Kolkata, IIT Kharagpur and Jadavpur University
Gujarati	DDU Nadiad, Gujrat
Bodo	Guhati University, Assam
Hindi	IIT Bombay
Kannad	Amrita University, Koimbatore
Kashmiri	Kashmir University, Srinagar
Malayalam	Amrita University, Koimbatore
Manipuri	Manipur University, Imphal Manipur
Marathi	IIT Bombay
Nepali	Assam University, Silchar Assam
Oriya	University of hydrabad
Punjabi	Thapar University and Punjabi university, Patiala
Sanskrit	IIT Bombay
Tamil	Tamil University, Thanjavur and Amrita University
Telugu	University of Hydrabad and Dravidian University, Kuppam
Urdu	University of Hydrabad and International Institute of Information Technology, Allahabad

Figure 1.1, shows the languages which are developing their WordNet under the IndoWordNet project. Expansion approach is being used to develop WordNets for these languages.

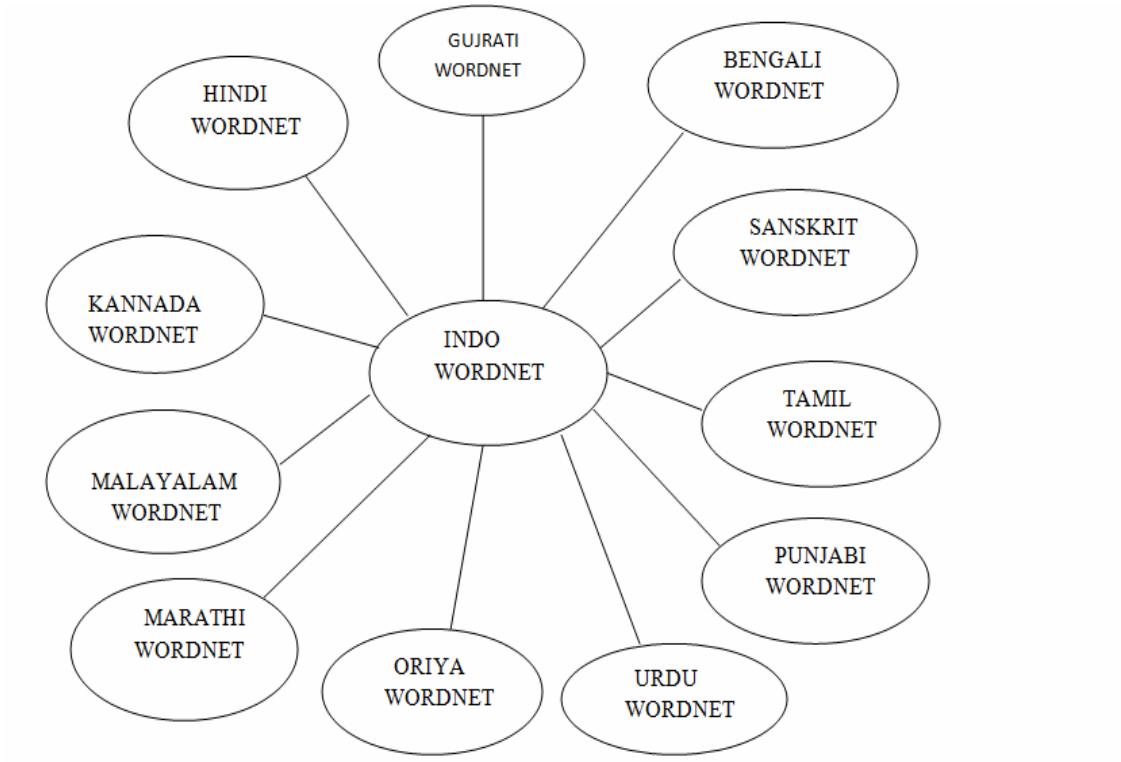


Figure 1.1: Linked IndoWordNet structure [7]

1.5 Differences between IndoWordNet (IWN) and EuroWordNet (EWN)

The expansion approach of WordNet creation adopted in EuroWordNet is also the principal methodology used in case of IndWordNet construction. In EWN, English provided the Interlingual Index (ILI). In IWN, the same is provided by Hindi. But there are some important differences between IWN and EWN:

- Right from the beginning, IWN insisted on storing lexical links expressing relationship of derivational morphology. Indian languages are rich in morphology. In Sanskrit WordNet, For example, the theory that all words are derived from verbal roots- '*dhaatus*'- is being seriously examined for its use as a fundamental guiding principle for storing and linking words.

- Causative verb forms are a typically occurring phenomenon in Indian languages. For example, □□□□ *(to eat),* □□□□□□ *(to feed)* and □□□□□□□ *(to cause to feed)* are forms derived from the same root □□□□ [8]. It has been decided to take special care to store causative forms in IWN and link them to their basic roots.

- Complex predicates also known as complex verbs are found in South Asian languages. They occur in the form of nominal+verb combinations called as conjunct verbs and verb+verb combinations called as compound verbs. IWN is drawing heavily on the research on complex predicates and is devising means for storing them [8].
- IWN has from the start taken part of speech linkages very seriously, especially between nouns and verbs. Ability and capability links between nouns and verbs are being incorporated exhaustively [6].
- IWN has finer categories for antonymy and gradation relations compared to EWN.

Chapter 2
Literature Review

2.1 WordNet Principle

WordNets have emerged as crucial resources for Natural Language Processing (NLP). WordNet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory [5]. A WordNet for a language is a linked structure of concept nodes represented by sets of synonymous words called *synsets*, which are connected through lexico-semantic relations. For the user, the WordNet is a rich lexicon like database that is queried using a browser to obtain information about words. Each word meaning can be represented by a set of word-forms known as *synonym set*s or *synset*s. Synsets are created for content words, *i.e.*, for Noun, Verb, Adjective and Adverb. In WordNet design, the focus will shift from words to concepts. For example, □□□□ (Sun), □□□□ (Earth), □□/□□□□ (Water) *etc.* are very common concepts. After selecting a concept, all the words representing that concept become members of the set of synonymous words. The first WordNet in the world was built for English at Princeton University. Then followed WordNets for European Languages: EuroWordNet. Similarly, IndoWordNet is being developed for major Indian languages. WordNets for different Indian languages are being built following expansion approach using Hindi WordNet which has been developed at Indian Institute of Technology, Bombay (IITB).

2.1.1 Lexical Matrix

The basic idea of a WordNet can be presented through the lexical matrix. The Lexical Matrix is an abstract representation of the organization of lexical information. Word forms are imagined to be listed as headings for the columns and word meanings as headings for the rows. Rows express synonymy while columns express polysemy as shown in Table 2.1 [6]. An entry in a cell of the matrix implies that the Word form in that column can be used in an appropriate context to express the meaning in that row. Thus, entry $E_{1.1}$ implies that word form F_1 can be used to express word meaning M_1. If there are two entries in the same column, the word form is polysemous; if there are two entries in the same row, the two word forms are synonyms (relative to a context).

Table 2.1: Illustrating the concept of Lexical Matrix [6]

Word	**Word Forms**

meanings	F_1	F_2	F_3	……	F_n
M_1	$E_{1.1}$	$E_{1.2}$			
M_2		$E_{2.2}$			
M_3			$E_{3.3}$		
M_4					
……..				……..	
M_m					$E_{m.n}$

Mappings between forms and meanings are many: many—some forms have several different meanings, and some meanings can be expressed by several different forms. For example, the matrix as shown in Table 2.2, describes the mapping between word forms and meanings of the Punjabi word ਜੱਗ.

Table 2.2: Example Lexical Matrix [7]

Polysemous word in the Column F2

Word Meanings	**Word Forms**		
	F1	**F2**	**F3**
World	□□□□□ ਸੰਸਾਰ	Synonyms in the row	
Container	ਜੱਗ		
Party or Feast	ਜੱਗ		

2.1.2 Synset making

Synsets are the building blocks for a WordNet. There are three baisc principles of minimality, coverage and replaceability which govern the creation of the synsets and these are explained below:

- **Minimality**: Only the minimal set that can uniquely identify the concept is used to create the synset, *e.g.*, to denote the concept of '*room*' the synset is

{ਘਰ, ਕਮਰਾ} (*room*). The Punjabi word ਘਰ is ambiguous and cannot uniquely denote the concept of a '*room*' by itself. For example, it could also mean ਘਰ (*house)*, ਦੇਸ਼ (*native country)*, or ਪਰਿਵਾਰ (family*)*. The addition of ਕਮਰਾ (also meaning *room*) to the synset brings out this unique sense of '*room*' [8].

- **Coverage**: The synset should contain all the possible words which denote a concept. The words are listed in order of decreasing frequency of their occurrence in the corpus. *e.g.,* {ਘਰ, ਕਮਰਾ} (*room*) [8].

- **Replaceability**: The words forming the synset should be mutually replaceable in a specific context. Two synonyms may mutually replace each other in a context C, if the substitution of the one for the other in C does not alter the meaning of the sentence [8]. For example, {ਸਵਦੇਸ਼, ਘਰ} (*motherland*), these two words can be replaced to denote the concept of '*motherland*'.
  ਅਮੇਰਿਕਾ ਵਿੱਚ ਦੋ ਸਾਲ ਬਤੀਤ ਕਰਨ ਤੋਂ ਬਾਅਦ ਸ਼ਾਮ ਸਵਦੇਸ਼/ਘਰ ਵਾਪਸ ਆ ਗਿਆ.

  Literal translation:
  America in two years stay after Shyam motherland returned
  '*Shyam returned to his motherland after spending two years in America*'

2.2 General methodology for WordNet creation

There are two approaches which can be followed for WordNet construction and these are merge approach and the expansion approach.

- **Merge approach**: In this approach, different senses in which a word can be used is first recorded [8]. Then the lexicographers construct a synset for each of the sense, following the three principles of minimality, coverage and replaceability for synset creation.

- **Expansion approach**: In the expansion approach, the synsets of the WordNet of a given source language SL are provided. Each synset is carefully studied for its

meaning. Then the words of the target language TL, representing that meaning are collected and put together in a set in frequency order [8].

2.2.1 Comparison of merge and expansion approaches for building WordNet

Both the merge and expansion approaches have their advantages and disadvantages. In the merge approach, there is no distracting influence of another language, which particularly will happen when the lexicographer will encounter cultural and regional concepts of the source language. The quality of the WordNet will be good, if the synset maker has a good knowledge of the language. But the disadvantage for this process is that it is very time consuming.

In case of the expansion approach, the whole WordNet building process becomes well guided in the sense of following the synsets of the source language. Also it has the advantage of being able to borrow the semantic relations of the given WordNet [8]. This can help in saving a huge amount of time. However, the lexicographer can be distracted by synsets which represent cultural and regional concepts. Also there is a problem in finding "*own concepts*" of the target language *i.e*., the concepts which are found in target language only and are not used in the source language.

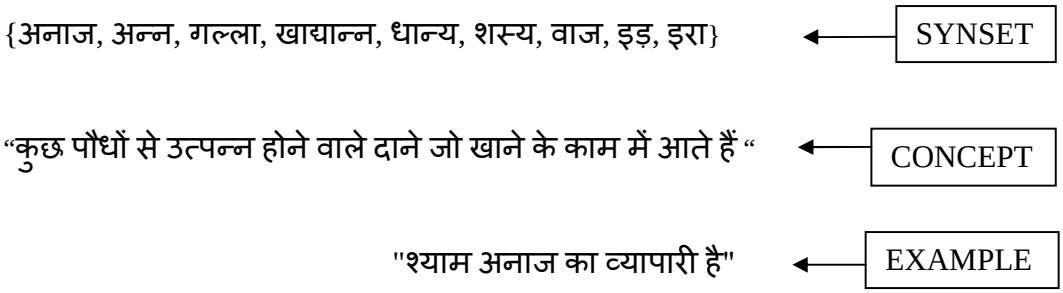
There is a predominance of the expansion approach in the WordNet building. This is primarily because many concepts are common across languages. Creating synsets for these universal concepts should be the first step in building a WordNet. If a language has already done this job, it is beneficial to take advantage from this work. Also semantic relations can be borrowed from the source language to be used in target language and it encourages to make use of the expansion approach in WordNet building process. If the source and target languages belong to the same language family, the expansion approach becomes more attractive, as distracting influences of cultural and regional concepts is minimal in this case.

2.3 Creation of Punjabi WordNet from Hindi WorNet using expansion approach

The WordNets follow the design principles of the English WordNet developed at Princeton University while paying particular attention to language specific phenomena

(such as complex predicates) whenever they arise. The Hindi WordNet(HWN), was developed by the researchers in the Centre for Indian Language Technology (CFILT), IIT Bombay, directed by Prof. Pushpak Bhattacharyya [9]. While HWN has been created by manually looking up the various listed meanings of words in different dictionaries, Punjabi WordNet (PWN) can be created from HWN by following expansion approach. That is, the synsets of HWN are modified to the synsets of PWN through addition or deletion of synonyms in the synset. Figure 2.1 shows the creation of the synset for the word □□□□ in PWN *via* addition and deletion of synonyms from HWN. The synset in HWN for this word is {अनाज, अन्न, गल्ला, खाद्यान्न, धान्य, शस्य, वाज, इड़, इरा}. PWN deletes {गल्ला, धान्य, शस्य, वाज, इड़, इरा} and adds {□□□} to it. Thus, the synset for □□□□ in PWN is {□□□□, □□□, □□□ □□□, □□□}. Hindi and Punjabi are close members of the same language family; so many Hindi words have the same meaning in Punjabi. This is especially for the words which are directly borrowed from Sanskrit. The semantic relations can be transferred directly, thus saving both time and effort.

Hindi WordNet entry



Corresponding Punjabi WordNet entry

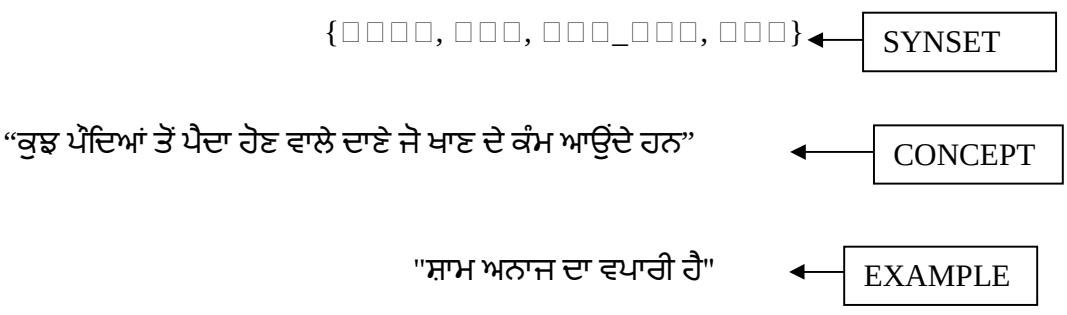


Figure 2.1: PWN synset creation from HWN [8]

2.4 Constituents elements of WordNet

WordNet is like a dictionary in that it stores words and meanings. However it differs from traditional dictionaries in many ways. For example, words in WordNet are arranged semantically instead of alphabetically. WordNet has nouns, verbs, adjectives, and adverbs arranged in synsets. Prepositions and conjunctions do not belong to any synset. Also, WordNet provides position of a word in ontology as an additional feature. Ontology is a hierarchical organization of concepts. For each category namely noun, verb, adjective and adverb, a separate ontological hierarchy is present. Each synset is mapped into some place in the ontology.

2.4.1 Synset and Concept

Synonymous words are grouped together to form synonym sets, or synsets. Each such synset therefore represents a single distinct sense or concept. Thus, the synset {□□□□□, □□□□□ □□□□, □□□□□□, □□□□ □□□□ , □□□□} will represent the sense as given in (2.1). {□□□ □□ □□□ □□□□ □□ □□□□ □□□□: □□□ □□□□□ □□□□ □□□ □□□ □□ □□□□□ □□ □□□□ □□□}.....(2.1)

A word sense is a particular meaning of a word. For example, the word ਨਿਕਲਣਾ has several meanings; as a noun it has two senses as shown in (2.2) and (2.3).

ਕਿਸੇ ਚੀਜ਼ ਦਾ ਆਪਣੀ ਥਾਂ ਤੋਂ ਆਉਣਾ ਜਾ ਵਿਖਾਯੀ ਦੇਣਾ .....(2.2)

ਬਤੀਤ ਕਰਣਾ .....(2.3)

A synset contains one or more synonymous word senses. For example, for the sense of the noun ਨਿਕਲਣਾ given in (2.2), the corresponding synset is given in (2.4).

{ ਨਿਕਲਣਾ , ਉਗਣਾ , ਚੜਣਾ } .....(2.4)

The synset is the fundamental organizational unit in WordNet. If a word can have multiple senses, it will appear in more than one synset. Each synset has a gloss (definition) associated with it. The gloss for the synset in (2.4) is given in (2.5).

“ਕਿਸੇ ਚੀਜ਼ ਦਿ ਦਾ ਆਪਣੀ ਥਾ ਤੋਂ ਆਉਣਾ ਜਾ ਵਿਖਾਯੀ ਦੇਣਾ ”…….. ...(2.5)

The synsets also have an example in addition to the gloss. For example, the gloss for the synset in (2.4) is given in (2.6).

“ਸੂਰਜ ਪੁਰਬ ਤੋਂ ਨਿਕਲਦਾ ਹੈ”……..….. (2.6)

The sense numbers in WordNet are assigned according to the frequency with which the word sense occurs in the corpus *i.e.*, the first sense of a word is usually more common than the second.

Different kinds of words found in WordNet are:

- **Polysemous:** Words which have multiple senses are known as Polysemous. In WordNet, each word occurs in as many synsets as it has senses. For example, the word □□□□□□ occurs in two noun synsets, { □□□ □□□□, □□□□ □□□□, □□□□□, □□□□, □□□□□□ } and { □□□□□□, □□□□□□, ਵੱਧਣਾ }.

- **Monosemous:** Words which can have only one sense are known as monosemous. For example, the word □□□□□ has only one sense and hence it will appear in only one synset.

- **Compound Words:** Besides single words, WordNet synsets also sometimes contain compound words which are made up of two or more words but are treated like single words in all respects.For example, WordNet has two–word compounds like □□□□□ □□□□ and □□□□ □□□□, three–word compounds like □□□□□□ □□□ □□□□ *etc.*

2.4.2 Relations in WordNet

The basic relations in WordNet are Semantic relations and Lexical relations which are explained as below:

2.4.2.1 Semantic Relations

Semantic relations are the relations between two whole synsets. Semantic relations are reciprocated *i.e.*, if there is a semantic relation '*R*' between meaning {*x*, *x*¢, . . } and meaning {*y*, *y*¢, . . .}, then there is also a relation '*R*¢' between {*y*, *y*¢, .. .} and {*x*, *x*¢, . . .}.Table 2.3 shows different kinds of semantic relations present between two synsets.

Table 2.3: Semantic Relations in WordNet [6]

Relation	Meaning
Synonymy	Similarity of meaning
Hypernymy/Hyponymy	Is-A (Kind-Of)
Entailment/Troponymy	Manner-Of (for verbs)
Meronymy/Holonymy	Has-A (Part-Whole)

- **Synonymy:** Synonymy means similarity of meaning. This relation is used to represent the words that have similar meanings. The relation is symmetric: if '*x*' is similar to '*y*', then '*y*' is equally similar to '*x*' [5]. Following words represent the synonymy relation between the words. For example, the word □□□□□□ (freedom) has synset{ ਸੁਤੰਤਰਤਾ , ਖ਼ਲਾੱਸੀ, ਛੁੱਟੀ , ਨਿਜਾਤ }. Similarly, synsets for different synonymous words are shown in table 2.4.

Table 2.4: Examples of Synonyms [11]

ਤੇਜ਼	ਮਾਹਿਰ, ਕੁਸ਼ਲ, ਤੇਜ਼, ਪਰਪੱਕ, ਨਿਪੁੰਨ, ਪ੍ਰਵੀਨ, ਪਰਬੀਨ, ਪਰਵੀਨ, ਹੁਸ਼ਿਆਰ, ਤਜਰਬੇਕਾਰ, ਉਸਤਾਦ
ਸ਼ਰੀਫ	ਸੱਭਿਅ, ਸੱਭਿਅਕ, ਭੱਦਰ, ਸ਼ਰੀਫ, ਸ਼ੁਸੀਲ, ਸਾਊ, ਭਲਾ, ਨੇਕ, ਸਲੀਕੇਮੰਦ, ਆਚਾਰਵਾਨ, ਸਮਾਜਿਕ
ਅਨਜਾਣ	ਅਨਜਾਣ, ਅਪੱਕ, ਅਪ੍ਰਪੱਕ, ਨੈਸਿੱਖਿਆ, ਨੈਸਿਖੂਆ, ਕੱਚ_ਘੜ, ਅਸਿੱਧ, ਕੱਚਾ
ਧਾਰਮਿਕਸਥਾਨ	ਪਵਿੱਤਰ-ਸਥਾਨ, ਧਾਰਮਿਕਸਥਾਨ, ਪਵਿੱਤਰ-ਅਸਥਾਨ, ਪਵਿੱਤਰ-ਥਾਂ, ਪੁੰਨ-ਭੂਮੀ, ਪਾਵਨ_ਭੂਮੀ

- **Hypernymy/Hyponymy:** Synsets are organized in a hierarchy via super-class/sub-class relationship (referred to as hypernymy/hyponymy) [5]. Hyponymy / hypernymy is a semantic relation between word meanings. For example, { ਪਿੱਪਲ } is a hyponym of ਪੇੜ and ਪੇੜ is a hyponym of {ਬੂਟਾ }.This relation is called hyponymy/hypernymy (variously called subordination / superordination, subset/ superset, or the ISA relation). A hyponym inherits all the features of a more generic concept and adds at least one feature that distinguishes it from its superordinate and from any other hyponyms of that superordinate. For example, ਪਿੱਪਲ inherits the features of its superordinate, ਪੇੜ (tree), but is distinguished from other trees by the hardness of its wood, the shape of its leaves, the use of its sap for syrup *etc*. This convention provides the central organizing principle for the nouns in WordNet [7].

  For example, □□□□□ (Pigeon) inherits the features from superordinate □□□□ (bird), but is distinguished from other □□□□ (bird) by color, size and living conditions as shown in Figure 2.2.

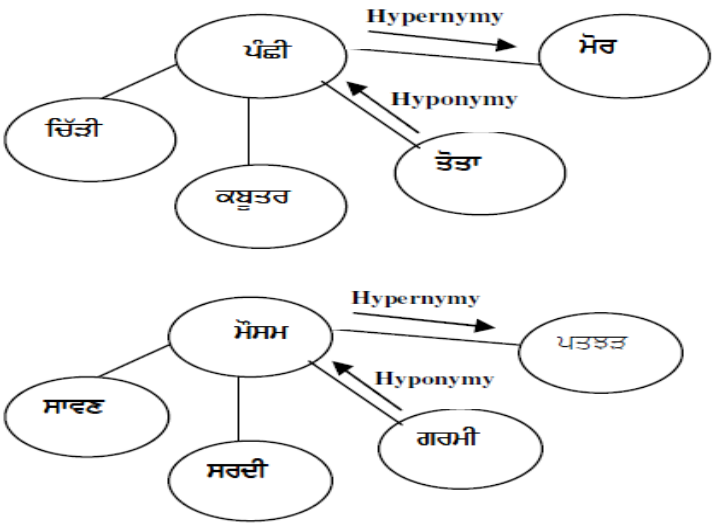


Figure 2.2: Hypernymy/Hyponymy relations [10]

- **Meronymy/Holonymy**: Another semantic relation—is the part-whole (or HAS A) relation, known as Meronymy /Holonymy [5]. For example, ਅੱਖਾਂ (eyes), ਬਾਂਹ (arm) and ਸਿਰ (head) are all parts of ਸ਼ਰੀਰ (body). This represents the Meronymy/Holonymy relation. ਸ਼ਰੀਰ (body) has a ਸਿਰ (head). ਸ਼ਰੀਰ (body) – Meronym and ਸਿਰ (head) -Holonym. The different examples depicting Meronymy and Holonymy relations are shown in Figure 2.3.

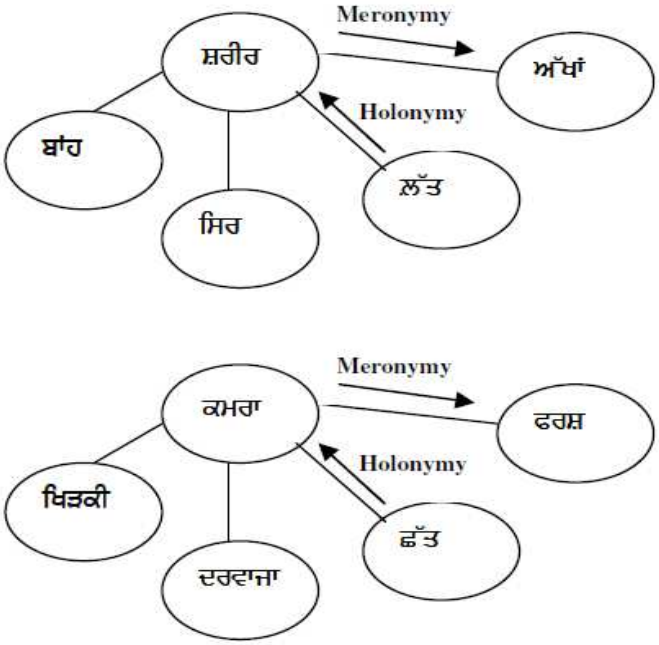


Figure 2.3: Meronymy/Holonymy relations [10]

- **Troponymy**: Troponymy is a semantic relation between two verbs when one is a specific '*manner*' elaboration of another [9]. For verbs, the term troponym is used instead of Hyponym. A troponym is a way of doing something else. For example, { ਉੱਡੀ ਫਿਰਨਾ }(soar) is a troponym of { ਉੱਡਣਾ , ਖੰਬ , ਪੰਖ , ਪਰ }(fly) because soaring is a way of flying. WordNet still uses the term Hypernym as the inverse of troponym; therefore { ਉੱਡਣਾ , ਖੰਬ , ਪੰਖ , ਪਰ }(fly) is a Hypernym of { ਉੱਡੀ ਫਿਰਨਾ }.

- **Entailment**: Entailment is a semantic relationship between two verbs. A verb '*A*' entails a verb '*B*', if the meaning of '*B*' follows logically and is strictly included in the meaning of '*A*'. This relation is unidirectional [9]. For example, the verb synset **{**ਚੱਲਣਾ, ਤੁਰਨਾ, ਪੈਦਲ ਚੱਲਣਾ**}** has an entailment relationship with the ਕਦਮ ਚੁੱਕਣਾ meaning of the verb synset{ ਕਦਮ , ਪੈਰ }, since walking entails stepping.

2.4.2.2 Lexical Relations

Lexical relations are the relations between members of two different synsets. The difference between lexical and semantic relations is that lexical relations are relations between members of two different synsets, but semantic relations are relations between two whole synsets [5].

- **Antonymy:** Antonymy is a lexical relation which turns out to be surprisingly difficult to define. The antonym of a word '*x*' is sometimes '*not-x*', but not always. For example, '*rich*' and '*poor*' are antonyms, but to say that someone is not rich does not imply that they must be poor; many people consider themselves neither rich nor poor. Antonymy [5], which seems to be a simple symmetric relation, is actually quite complex. Antonymy is a lexical relation between word forms, not a semantic relation between word meanings. For example, the word ਨੇੜੇ (near) has the antonym as ਦੂਰ (far), the word ਮੋਟਾ (fat) has the antonym as ਪਤਲਾ (thin).

- **Gradation:** This lexical relation will provide possible intermediate state between two antonyms. For example, to show gradation relation among time words we have, □□□□□ '*noon*' between {ਸਵੇਰ} '*morning*' and { ਸ਼ਾਮ } '*evening*' [9].

2.5 Demonstration of all Relations in WordNet

Figure 2.4 shows all the relations like synonymy, hypernymy, hyponymy, meronymy *etc.* for synset {□□ (home), □□□□□}. The hypernymy relation (Is-A) of it, is linked to { □□□□□ , □□□□□□ , □□□□ }.Its meronymy relation (Has-A) is linked to {□□□□□□,□□□□□ }and { □□□□ } and hyponymy relation to {□□□□, □□□□□□□}, {□□□□□□ , ਕੁੱਲੀ } and {□□□□□, □□□□□ } [6] .

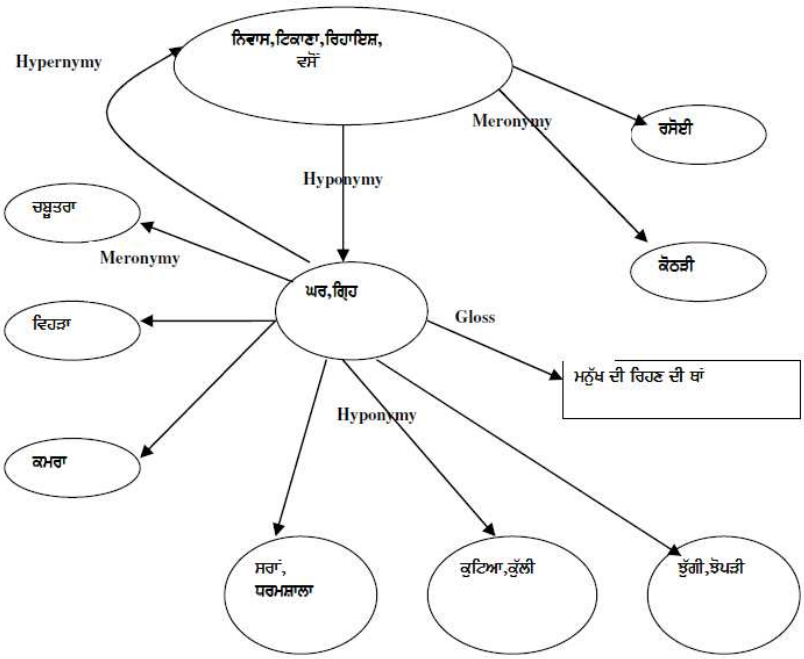


Figure 2.4: Relations for synset of □□ (Home) [6]

2.6 Existing electronic dictionaries in Punjabi

There are many dictionary tools available in Punjabi. Some of the paper dictionaries have been converted into electronic dictionaries, while some have been specially made as electronic dictionaries. Many online dictionaries have also been developed. Some of the popular electronic dictionaries in Punjabi are listed below:

- **Punjabi Kosh:** Punjabi Kosh is an English-to-Punjabi and Punjabi-to-English dictionary designed by Noah Hart. A user can make use of a keypad to enter a word and it allows dictionary use in both languages. It is a very useful tool for learning Punjabi. The dictionary is freely available at [12].

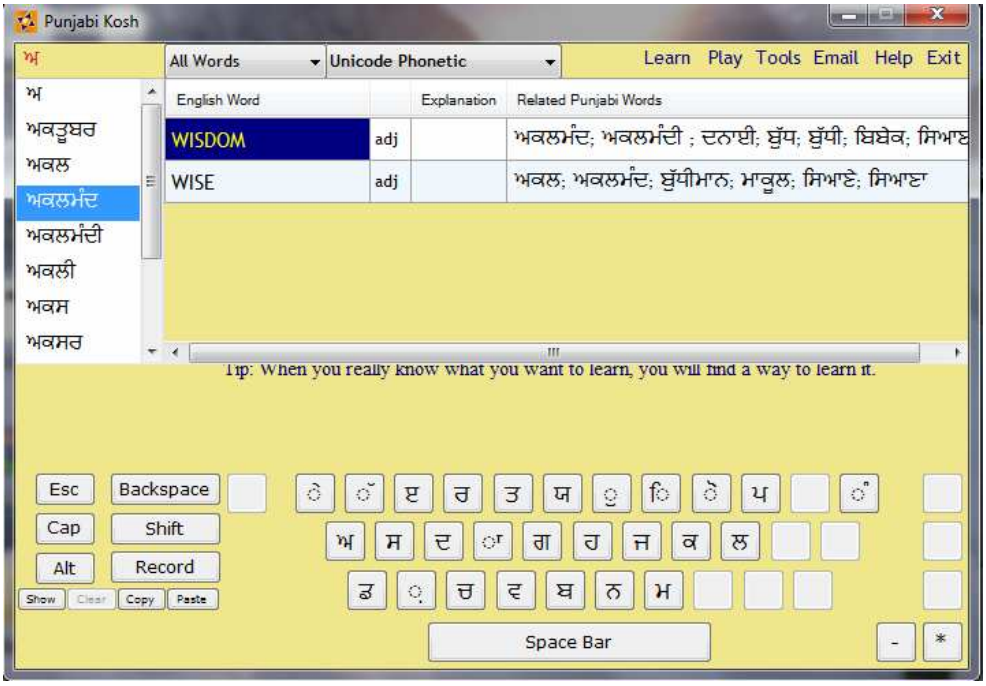


Figure 2.5: Web interface for Punjabi Kosh [12]

- **Punjabi Shabdkosh:** Punjabi Shabdkosh is a Punjabi to English dictionary and it has been developed by Harwinder Singh Tiwana. The dictionary is freely available at [13].

- **Punjabi Dictionary by CDAC:** An ISCII based Punjabi-English dictionary developed by CDAC is made available on the language CD freely made available by MCIT. The GIST typing tools have to be used for typing and searching for words in Punjabi.

- **Punjabi-English Dictionary:** The Punjabi-English paper dictionary developed by Punjabi University, Patiala has been converted into electronic form. The electronic dictionary displays the Punjabi words both in Gurmukhi and Shahmukhi scripts. This dictionary has about 31,000 entries [14].

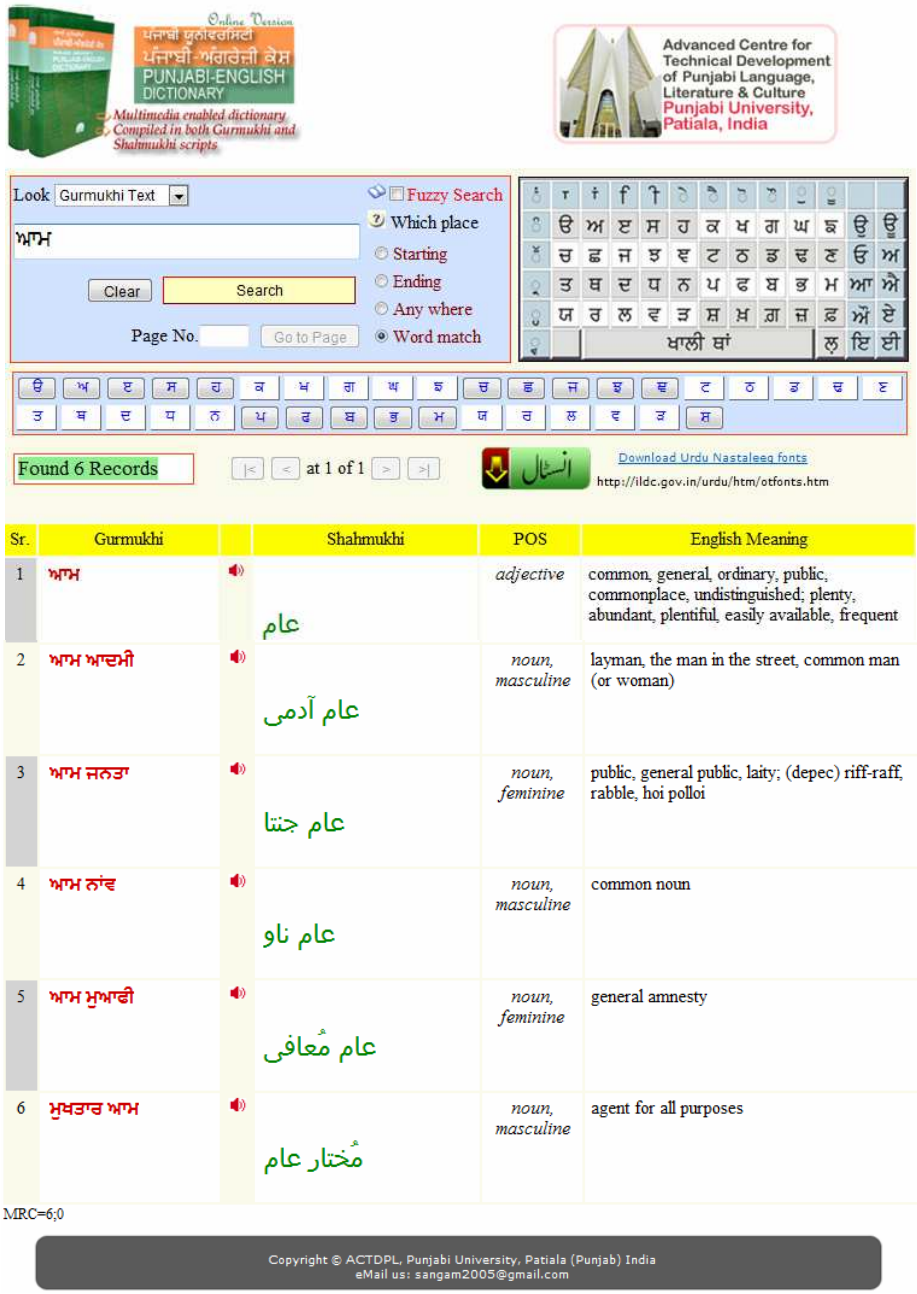


Figure 2.6: Web interface for Punjabi English dictionary [14]

- **Punjabi-English and English-Punjabi Dictionary by Jasjit Singh Thind** : An online Punjabi-English and English-Punjabi dictionary has been provided by Jasjit Singh Thind [15]. This dictionary is available on the website www.punjabionline.com.

Figure 2.7: Web interface showing different senses for word "*life*" in Punjabi [15]

- **Punjabi Encyclopedia and Gurbani Dictionary by Dr. Kulbir Singh Thind:** A powerful online search facility for simultaneously searching for any Punjabi word in Mahan Kosh Encyclopedia, Gurbani Dictionaries and Punjabi/English Dictionaries is provided on the website www.srigranth.org by Dr. Kulbir Singh Thind [16].

Figure 2.8: Web interface showing different senses for word "*health*" [16]

- **Shabdkosh :** This dictionary was implemented by Maneesh Soni in 2003. The word 'Shabdkosh' is the Romanized English spelling of the Hindi word for 'Dictionary'. The online English Punjabi dictionary started as a small project with an aim to just make a website in Punjabi [17]. The main reason to choose Dictionary as an application was unavailability of a good quality online dictionary at the time. This dictionary provides different senses of a word along with definition and its synonymous words. This site now provides dictionaries in other language as well like Hindi, Bengali and Gujarati *etc.* This dictionary has grown in popularity over the years because of its easy interface and strong vocabulary database.

Figure 2.9: Web interface for Shabdkosh [17]

Chapter-3

Problem Statement

3.1 Motivation

People throughout the world have been using computers and Internet in their own languages. So far, Indian users in general and Punjabi users in particular have been compelled to use them in English despite the dominance of Indian engineers and scientists in the Information Technology world. Unless we support our own languages on the technological front, it is impossible to use IT or internet to uplift and improve the socio-economic condition of our country. There is a need for language based content and technology. The society at large can benefit from the Information Technology effectively if people can communicate with computers in their own languages. Barely 65 % of our population is literate, of which only an elite minority (~10%) can read, write, and speak the English language. This shuts out most of the Indian population from the worldwide web and its huge potential [18]. As more and more Punjabi texts can be accessed in electronic form and one can see a number of web pages appearing on a daily basis in Punjabi language, there is a great need to create large scale online lexical-semantic nets for Punjabi so as to be used in Natural Language Processing, Information Retrieval, and other areas.

WordNet is a very rich resource of lexical information. WordNet has proved very useful for different activities in Natural Language Processing, *e.g.*, parsing and machine translation, concept identification, Word Sense Disambiguation, the treatment of syntactic and semantic ambiguity *etc.* The first WordNet was developed for English at Princeton University. The success of English WordNet has inspired several projects that aim at constructing WordNets for other languages. Hindi WordNet was the first to be

developed for any of the Indian languages and in the year 2006 it was made available free for research. It is therefore inspirational to build Punjabi WordNet.

3.2 Gap Analysis

WordNet is available for different Indian languages like Hindi, Marathi, Gujrati, Bengali,Telugu *etc.* but no relevant work has been done in case of Punjabi WordNet and is unavailable upto now. Punjabi WordNet is now being developed at Thapar University under the IndoWordNet project of IIT, Bombay. Punjabi WordNet is being developed from Hindi WordNet by following the expansion approach.

Most of the online dictionaries available on the web for Punjabi language have used English Punjabi as the language pair and a very little work has been done taking Hindi Punjabi as the language pair. Moreover, the dictionaries presently available provide a simple translation of the given word in target language. These dictionaries do not provide additional information about the word like the concept of the word, an example sentence depicting the word usage, part of speech to which the word belongs, synonym words available for each of the sense of the word *etc*.

3.3 Problem statement

There is a need to create a Punjabi WordNet because large scale online lexical-semantic nets are required for the Punjabi language. Punjabi WordNet can be used in Natural Language Processing, Information Retrieval, and other related areas. Hindi WordNet can be used to build Punjabi WordNet by following the Expansion approach. Expansion approach is being used as both Hindi and Punjabi belong to the same language family and the Punjabi WordNet making process will become well guided and can save an enormous amount of time. This resource once developed can be used for creation of Hindi Punjabi bilingual dictionary. By using Punjabi WordNet we can enrich our bilingual dictionary as it will not simply provide meaning of the word in target language, but it will also explain

the concept of the word, category of word representing its part of speech, an example sentence and synonymous words. Many English Punjabi dictionaries are available on the web but very little work has been done for Hindi Punjabi dictionaries. The main aim of this thesis is to develop Punjabi WordNet and Punjabi Hindi bilingual dictionary and a web application so that these resources can be accessed online.

3.4 Objectives

The main objectives of this work are:

- To create Punjabi WordNet from Hindi WordNet using Expansion approach.
- Usage of Punjabi WordNet for creation of Hindi Punjabi and Punjabi Hindi bilingual dictionary with features like concept, an example sentence and synonymous words.
- To develop a web application for Punjabi WordNet and Punjabi Hindi bilingual dictionary.

3.5 Methodology

To achieve the objectives discussed in section 3.4, a step-by-step methodology has been followed. The detail of this is given below.

- Study of existing WordNets and their functionality has been carried out.
- Analysis of different WordNet relations like Synonymy, Antonymy, Hypernymy-Hyponymy, Holonymy-Meronymy, Entailment,Troponymy *etc.* has been performed.
- A tool known as IL-MultiDic Development tool provided by IIT, Bombay has been used for creating Punjabi WordNet.
- Hindi Punjabi bilingual dictionary has been created by using source and target files of IL-MultiDic Development tool.
- Web application for bilingual dictionary has been developed by using JSP at the front end and text files at the back end.

.Chapter-4

Design and implmentation of Punjabi WordNet and Punjabi Hindi billinual dictionary

In developing the Punjabi WordNet, it has been convenient to divide the work into two interdependent tasks:

- The first task has been to write the source files for Punjabi using the IL-MultiDic development tool which has been provided by IIT, Bombay.
- The second task has been to create a set of computer programs in Java Netbeans 7.0 platform that would make use of these source files at the back end to develop the Punjabi WordNet and Punjabi Hindi bilingual dictionary.

4.1 Creation of synsets for Punjabi using IL- MultiDic development tool

For creating the synsets in Punjabi language, a tool known as IL-MultiDic development tool has been used. This tool is provided by IIT, Bombay and it is designed to help in standardization of the lexical data entries by using the same Id numbers to represent similar concepts in source and target languages. The tool is being used to write the source files that would provide information of Punjabi words and this data will be used for developing Punjabi WordNet. The data values are to be generated for Punjabi in correspondence to the entries done for Hindi WordNet using this tool. This will help in standardizing the lexical entries to be used.

4.1.1 Configuration of tool

During configuration of the tool, source and target files are to be selected. The source and target files act as input and output to the tool. The source file consists of synsets which have already been created and which will guide the synset creation for target language. In this case, the source file consists of 7168 core synsets of Hindi language. The target file will consist of synsets which will be created using this tool. Browse button can be used to choose the source and target files. Target language will specify the language of the target file. In this case, target language will be Punjabi and Target font will be Arial Unicode MS as shown in Figure 4.1.

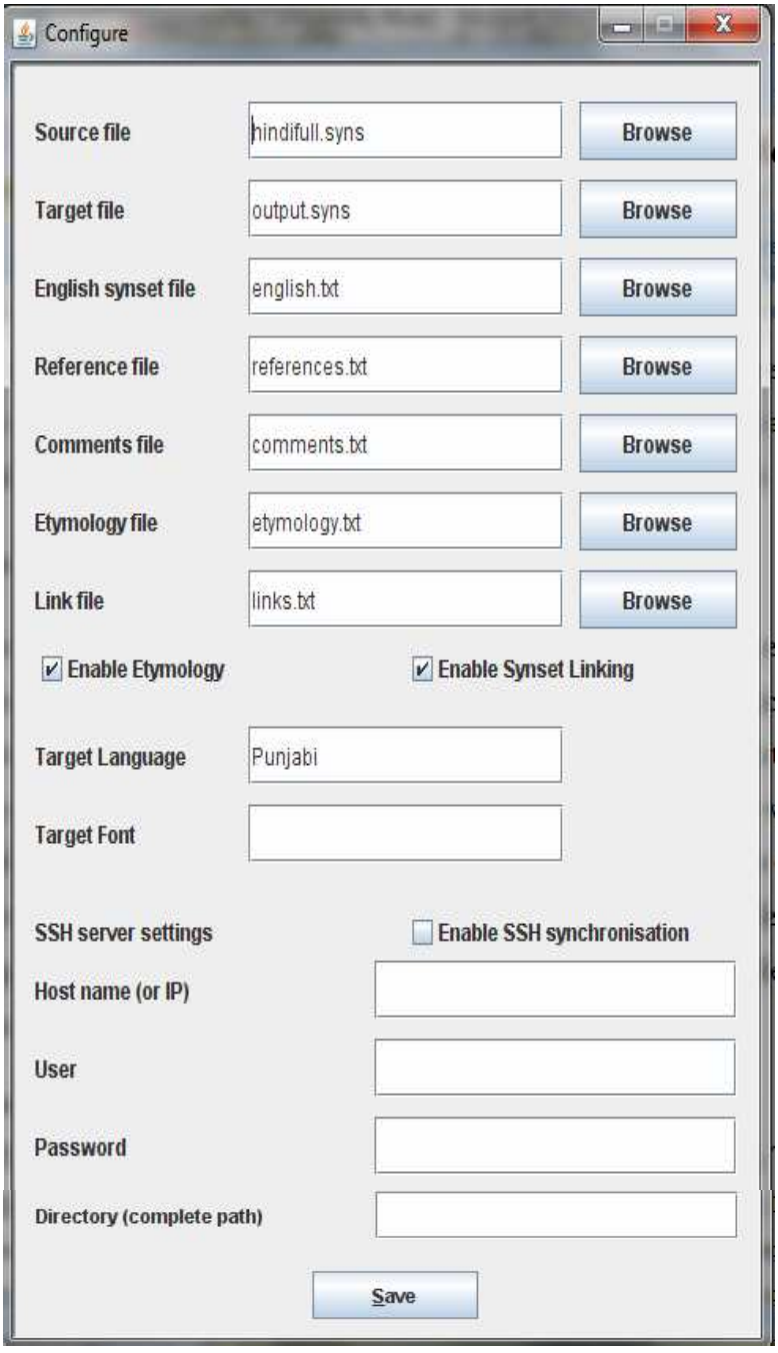


Figure 4.1: Configuring IL-MultiDic development tool

4.2.2 Working of the tool

This tool is a user friendly tool and it is used to link together the synsets that convey the similar meaning in different languages. In this case, the tool is being used to link the synsets for Hindi and Punjabi language.

- First of all source and target files are uploaded during configuration of the tool.
- The files which contain the data entries for Hindi words are the source files. There are source files of core synsets which contain around 7168 synsets. Each synset has been assigned a unique id number which will uniquely recognize the concept represented by that synset. Figure 4.2, shows the tool used for the development of Punjabi synsets corresponding to each Hindi word entry.

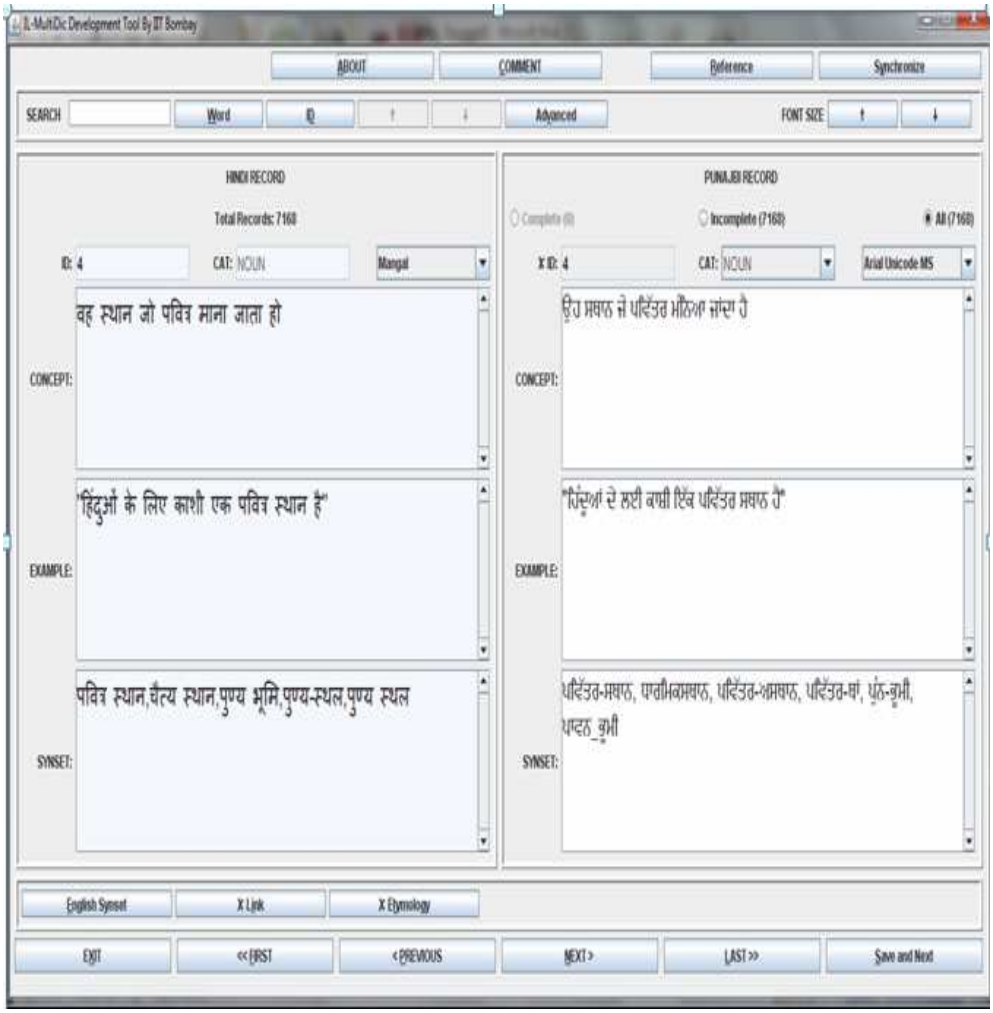


Figure 4.2: Tool for creating standardized lexical data

- The files used by Hindi WordNet which contains 7168 entries are uploaded on the left hand side. Each source file contains a list of synsets for all parts of speech, *i.e.*, noun, verb, adjective and adverb. Each of the word record contains Id number, synonymous word forms, category, concept and an example sentence. The source file for Hindi has been shown in the Figure 4.3 which has been used to generate the file for Punjabi having same standard id number indicating a single concept.

ID	106
CAT	Noun
CONCEPT	स्वीकार करने की क्रिया या भाव
EXAMPLE	भारत सरकार ने इस परियोजना को चालू करने के लिए अपनी स्वीकृति दे दी है"
SYNSET-HINDI	स्वीकृति, मंजूरी, इक़रार, इकरार, अंगीकरण, अंगीकृति, अनुज्ञप्ति, संप्रत्यय, ईजाब

ID	107
CAT	Noun
CONCEPT	अपने मन से यह समझने की क्रिया या भाव कि ऐसा हो सकता है या होगा
EXAMPLE	कभी-कभी अनुमान गलत भी हो जाता है
SYNSET-HINDI	अनुमान, अंदाज़, अंदाज, अंदाज़ा, अंदाजा, अन्दाज़, अन्दाज, अन्दाज़ा, अन्दाजा, अटकल, कूत, अटकर, अरसट्टा, अड़सट्टा, अनुमिति, तख़मीना, तखमीना

ID	108
CAT	Noun
CONCEPT	वह शक्ति या भाव जो मन में नयी, अनोखी, अनदेखी, अनसुनी आदि बातों के स्वरूप को उपस्थित करती है
EXAMPLE	मूर्तिकार की कल्पना पत्थर को तराश कर मूर्त रूप प्रदान करती है
SYNSET-HINDI	कल्पना, ख़याल, खयाल, ख्याल, फंतासी, कल्पना शक्ति, तसव्वुर, तसव्वर, तसौवर

ID	109
CAT	Noun
CONCEPT	किसी के किए हुए काम या सामने रखे हुए सुझाव को ठीक मानकर

	अपनी दी हुई स्वीकृति
EXAMPLE	हम इस प्रस्ताव का अनुमोदन करते हैं
SYNSET-HINDI	समर्थन, हिमायत, अनुमोदन, ताईद
ID	110
CAT	Noun
CONCEPT	किसी बात, सुझाव आदि पर प्रसन्नता प्रकट करने की क्रिया या भाव
EXAMPLE	मेरे वक्तव्य पर उसकी वाहवाही दिखावटी है
SYNSET-HINDI	वाहवाही, अनुमोदन, आमोदन

Figure 4.3: Source file of Hindi synsets

- For each data value for Hindi, corresponding entry for Punjabi is made on the right hand side. This entry is saved and then move on to next data value by a click on the Save and Next button. All the data entries are to be completed in each file following the same pattern.

- Finally the output of the tool is obtained which is the file containing Punjabi word entries. This output file will have the same format as that of the source file of Hindi words entries. Thus the Punjabi word file has been generated having the same standardized id numbers as that of the input Hindi file. The output of the tool is shown in Figure 4.4.

- Each entry in the output file includes a unique Id number for the word, the concept which it represents, an example sentence and a synset showing synonymous words for that concept. The information also includes a category field for each entry which will represent the part-of-speech, *i.e.*, noun, verb *etc.* to which the word belongs.

ID	106
CAT	noun
CONCEPT	ਸਵੀਕਾਰ ਕਰਨ ਦੀ ਕਿਰਿਆ ਜਾਂ ਭਾਵ
EXAMPLE	ਭਾਰਤ ਸਰਕਾਰ ਨੇ ਇਸ ਯੋਜਨਾ ਨੂੰ ਚਾਲੂ ਕਰਨ ਦੇ ਲਈ ਆਪਣੀ ਸਹਿਮਤੀ ਦੇ ਦਿੱਤੀ ਹੈ

SYNSET-PUNJABI	ਸਹਿਮਤੀ, ਮਨਜ਼ੂਰੀ, ਆਗਿਆ, ਇਜਾਜ਼ਤ, ਹੁਕਮ, ਇਕਰਾਰ, ਅਨੁਮਤੀ, ਪ੍ਰਵਾਨਗੀ
ID	107
CAT	noun
CONCEPT	ਆਪਣੇ ਮਨ ਤੋਂ ਇਹ ਸਮਝਣ ਦੀ ਕਿਰਿਆ ਜਾਂ ਭਾਵ ਕਿ ਅਜਿਹਾ ਹੋ ਸਕਦਾ ਹੈ ਜਾਂ ਹੋਵੇਗਾ
EXAMPLE	ਕਦੇ ਕਦੇ ਅੰਦਾਜਾ ਗ਼ਲਤ ਵੀ ਹੋ ਜਾਂਦਾ ਹੈ
SYNSET-PUNJABI	ਅੰਦਾਜਾ, ਅਨੁਮਾਨ, ਅਨੁਮਾਨਿਤ, ਤੁੱਕਾ, ਕਿਆਸ, ਅੰਦਾਜ਼ਾ, ਅਟਕਲ, ਅੱਟਾ-ਸੱਟਾ
ID	108
CAT	noun
CONCEPT	ਉਹ ਸ਼ਕਤੀ ਜਾਂ ਭਾਵ ਮਨ ਵਿਚ ਨਵੀ,ਅਨੇਖੀ,ਅਣਦੇਖੀ,ਅਣਸੁਣੀ ਆਦਿ ਗੱਲਾ ਦਾ ਸਰੂਪ ਉਜਾਗਰ ਕਰਦੀ ਹੈ
EXAMPLE	ਮੂਰਤੀਕਾਰ ਦੀ ਕਲਪਨਾ ਪੱਥਰ ਨੂੰ ਤਰਾਸ਼ ਕੇ ਮੂਰਤੀ ਦਾ ਰੂਪ ਪ੍ਰਦਾਨ ਕਰਦੀ ਹੈ
SYNSET-PUNJABI	ਕਲਪਨਾ, ਖਿਆਲ, ਕਲਪਨਾ-ਸ਼ਕਤੀ, ਅਨੁਮਾਨ, ਤਸੱਬੁਰ, ਮਨੋਵਿਰਤੀ, ਮਨੋਵ੍ਰਿਤੀ, ਮਨੋਬਿਰਤੀ, ਮਨੋ-ਖਿਆਲ
ID	109
CAT	Noun
CONCEPT	ਕਿਸੇ ਦੇ ਕੀਤੇ ਹੋਏ ਕੰਮ ਜਾਂ ਸਾਹਮਣੇ ਰੱਖੇ ਹੋਏ ਸੁਝਾਅ ਨੂੰ ਠੀਕ ਮੰਨ ਕੇ ਆਪਣੀ ਦਿੱਤੀ ਹੋਈ ਪ੍ਰਵਾਨਗੀ
EXAMPLE	ਅਸੀ ਇਸ ਪ੍ਰਸਤਾਵ ਦਾ ਸਮਰਥਨ ਕਰਦੇ ਹਾਂ
SYNSET-PUNJABI	ਰਜ਼ਾਮੰਦੀ, ਪ੍ਰਵਾਨਗੀ, ਸਮਰਥਨ, ਹਿਮਾਇਤ, ਤਾਈਦ, ਅਨੁਮੋਦਨ
ID	110
CAT	Noun
CONCEPT	ਕਿਸੇ ਗੱਲ,ਸੁਝਾਅ ਆਦਿ ਤੇ ਖੁਸ਼ੀ ਪ੍ਰਗਟ ਕਰਨ ਦੀ ਕਿਰਿਆ ਜਾਂ ਭਾਵ
EXAMPLE	ਮੇਰੇ ਬਿਆਨ ਤੇ ਉਸਦੀ ਪ੍ਰਸੰਸਾ ਪੂਰਵਕ ਟਿੱਪਣੀ ਦਿਖਾਵਟੀ ਹੈ
SYNSET-PUNJABI	ਟਿਪਣੀ, ਵਾਹ-ਵਾਹ

Figure 4.4: Output file containing entries for Punjabi words

4.2 Creation of Punjabi WordNet

The output file for Punjabi words as obtained from the IL-Multidic development tool is used as the back end to develop Punjabi WordNet. JSP is used at front end to develop the user interface for web application. The algorithm used in developing Punjabi WordNet is described in Algorithm 4.1.

Algorithm 4.1: Creation of Punjabi WordNet

1. Read the text file of Punjabi words and generate a linked list *L* for the file.
2. Input the word to be searched in a string *w*.
3. For each word *w* to be searched
   i. Search the linked list *L* for the occurrence of word *w* in synset.
   ii. If w is found in the list, then retrieve all the related ID numbers and store them in an array.
   iii. Search the linked list *L* for each of the ID number stored in array.
      a. If ID match occurs then retrieve the associated information with each ID number and store in a list *F*.
      b. Display the linked list *F*.

Processing of Punjabi synset file for browsing the Punjabi WordNet has been shown by taking an example word "ਵਸਤੂ". Search the occurrence of the word "ਵਸਤੂ" in the linked list *L*. The word is found in the list *L* and the corresponding ID numbers for each occurrence of the word will be retrieved from the list and stored in array. Array will have the following ID numbers: 744, 923 and 1500 as the word had three different occurrences in the list *L*. Search the list *L* for each of the ID number stored in array and retrieve the related information with each ID number from the list. Store this information in linked list *F*. The information associated with each of the ID number as contained in final list *F* is shown in the Figure 4.5.

ID	744
CAT	noun
CONCEPT	ਉਹ ਜਿਸਦਾ ਕੋਈ ਆਕਾਰ ਜਾਂ ਰੂਪ ਹੋਵੇ ਅਤੇ ਜੋ ਠੋਸ,ਸਰੀਰ ਆਦਿ ਦੇ ਰੂਪ ਵਿਚ ਹੋਵੇ

EXAMPLE	ਦੁੱਧ ਇਕ ਪੀਣ ਵਾਲਾ ਪਦਾਰਥ ਹੈ
SYNSET- PUNJABI	ਪਦਾਰਥ, ਵਸਤੂ, ਚੀਜ
ID	923
CAT	Noun
CONCEPT	ਵਾਸਤਵਿਕ ਜਾਂ ਕਲਪਿਤ ਸੱਤਾ
EXAMPLE	ਹਵਾ ਇਕ ਅਮੂਰਤ ਵਸਤੂ ਹੈ
SYNSET- PUNJABI	ਵਸਤੂ, ਚੀਜ, ਸ਼ੈ, ਚੀਜ਼
ID	1500
CAT	Noun
CONCEPT	ਉਹ ਵਸਤੂ ਜਿਸ ਦਾ ਕਿਸੇ ਕੰਮ ਵਿਚ ਉਪਯੋਗ ਹੁੰਦਾ ਹੈ
EXAMPLE	ਇੱਟਾਂ,ਸੀਮਿੰਟ ਆਦਿ ਸਾਮਾਨ ਘਰ ਬਣਾਉਣ ਦੇ ਕੰਮ ਆਉਂਦਾ ਹੈ
SYNSET-PUNJABI	ਸਮਾਨ, ਵਸਤੂ, ਚੀਜ, ਪਦਾਰਥ, ਸਮੱਗਰੀ

Figure 4.5: Final result list for the word “ਵਸਤੂ”

The Algorithm 4.1 is explained using the flow chart. Figure 4.6, shows the flowchart for Punjabi WordNet. First of all, a linked list *L* is generated for the text file of Punjabi words. When the user enters the word to be searched, the word is searched in the linked list *L*. If the word is found in the list, then the corresponding ID numbers for each occurrence of the word are retrieved and stored in an array. Again a search operation is performed on the linked list *L* for each of the ID number stored in array. If the ID match occurs, the corresponding information *i.e.,* the concept, category, example and synonyms of each ID number is fetched and this information is stored in a linked list *F*. Finally this linked list *F* is displayed.

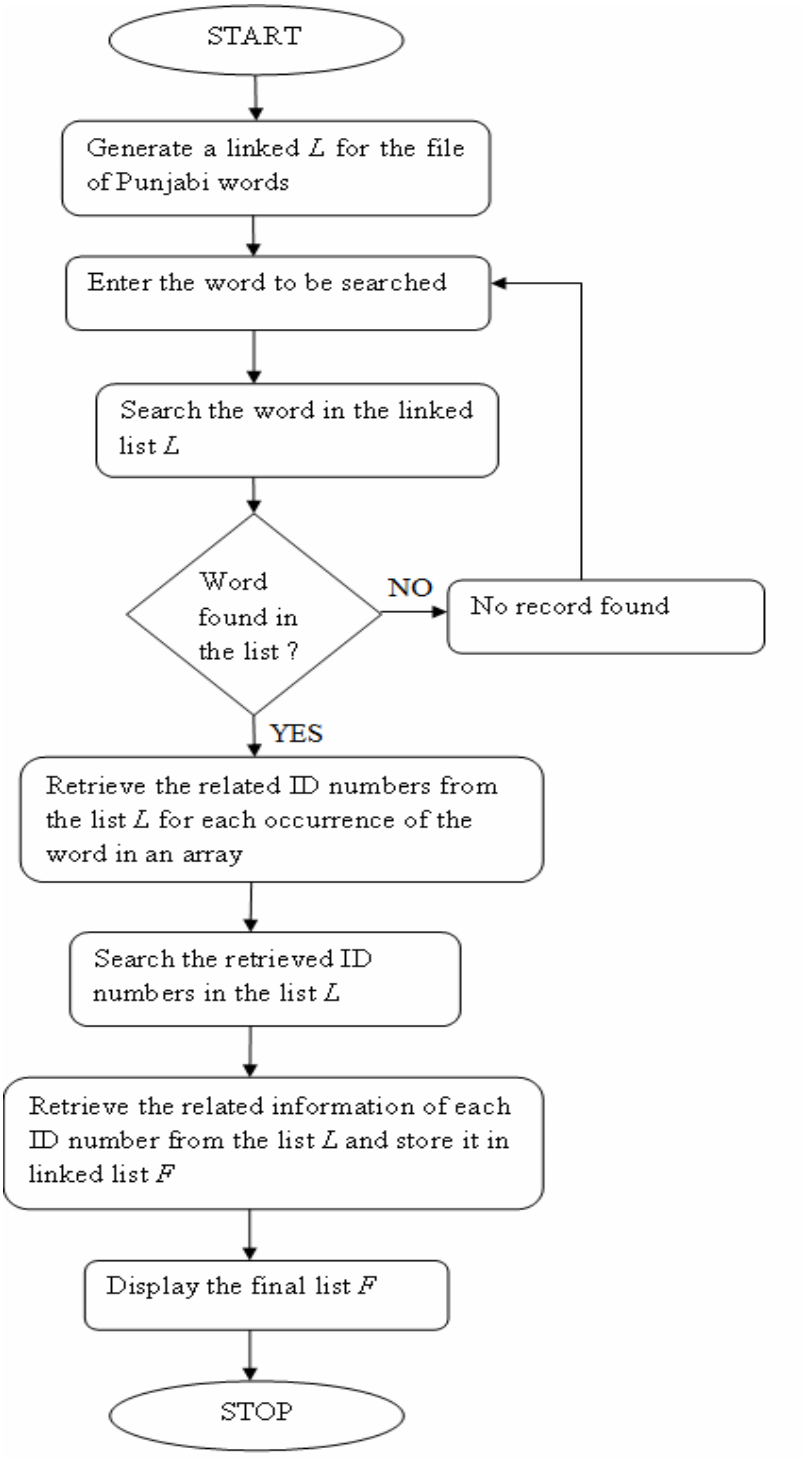


Figure 4.6: Flowchart for Punjabi WordNet

4.3 Creation of Hindi to Punjabi Dictionary

Hindi Punjabi dictionary has around 7168 entries. Each word in the dictionary has an associated information *i.e.,* category like noun, verb *etc*., concept which explains the word, example usage of the word, and synset and its equivalent in the Punjabi language.

Algorithm 4.2: Creation of Hindi to Punabi dictionary

1. Read the text file of Punjabi words and generate a linked list L_1 for the file. Similarly read the text file of Hindi words and generate a linked list L_2 for the file.
2. Input the word to be searched in a string *w*.
3. For each word *w* to be searched
   i. Search the linked list L_1 for the occurrence of word *w* in synset.
   ii. If w is found in the list, then retrieve all the related ID numbers and store them in an array.
   iii. Search the linked list L_2 for each of the ID number stored in array.
      a. If ID match occurs, then retrieve the associated information with each ID number and store in a linked list *F*.
      b. Display the linked list *F*.

Processing of Punjabi synset file for browsing the Punjabi WordNet has been shown by taking an example word "आम". Search for the occurrence of the word "आम" in the linked list L_1. The word is found in the list L_1 and the corresponding ID numbers for each occurrence of the word will be retrieved from the list and stored in array. Array will have the following ID numbers: 3462, 3463, 3468, 3469 as the word had four different occurrences in the list L_1. Search the list L_2 for each of the ID number stored in array and retrieve the related information with each ID number from the list. Store this information in linked list *F*. The information associated with each of the ID number as contained in final list *F* is shown in the Figure 4.7.

ID	3462
CAT	Noun
CONCEPT	ਇਕ ਫਲ ਜੋ ਖਾਇਆ ਜਾਂ ਚੂਸਿਆ ਜਾਂਦਾ ਹੈ
EXAMPLE	ਤੋਤਾ ਦਰੱਖਤ ਤੇ ਬੈਠ ਕੇ ਅੰਬ ਖਾ ਰਿਹਾ ਹੈ / ਸ਼ਾਸਤਰਾਂ ਨੇ ਅੰਬ ਨੂੰ

	ਇੰਦਰਾਸਨੀ ਫਲ ਦਾ ਨਾਮ ਦਿੱਤਾ ਹੈ
SYNSET- PUNJABI	ਅੰਬ, ਆਮ
ID	3463
CAT	Noun
CONCEPT	ਗਰਮ ਦੇਸ਼ਾ ਵਿਚ ਪਾਇਆ ਜਾਣ ਵਾਲਾ ਇਕ ਵੱਡਾ ਸਦਾਬਹਾਰ ਦਰੱਖਤ ਜਿਸਦੇ ਰਸੀਲੇ ਫਲ ਖਾਏ ਜਾਂ ਚੂਸੇ ਜਾਂਦੇ ਹਨ
EXAMPLE	ਅੰਬ ਦੀ ਲੱਕੜੀ ਦਾ ਉਪਯੋਗ ਸਜਾਵਟ ਦੀਆਂ ਵਸਤੂਆਂ ਬਣਾਉਣ ਵਿਚ ਕੀਤਾ ਜਾਂਦਾ ਹੈ
SYNSET- PUNJABI	ਅੰਬ, ਅੰਬ ਦਾ ਦਰੱਖਤ
ID	3468
CAT	Adjective
CONCEPT	ਜਿਸ ਵਿਚ ਕੋਈ ਵਿਸ਼ੇਸ਼ਤਾ ਨਾ ਹੋਵੇ ਜਾਂ ਵਧੀਆ ਤੋਂ ਘੱਟ ਦਰਜੇ ਦਾ
EXAMPLE	ਇਹ ਸਧਾਰਨ ਸਾੜੀ ਹੈ / ਇਹ ਕੰਮਚਲਾਉ ਸਰਕਾਰ ਬਹੁਤੇ ਦਿਨਾਂ ਤੱਕ ਨਹੀਂ ਟਿਕਣ ਵਾਲੀ
SYNSET- PUNJABI	ਸਧਾਰਨ, ਆਮ, ਕੰਮ_ਚਲਾਉ, ਮਾਮੂਲੀ
ID	3469
CAT	Adjective
CONCEPT	ਜਿਆਦਾਤਰ ਸਭ ਵਿਅਕਤੀਆਂ,ਮੋਕਿਆਂ,ਅਵਸਥਾਵਾਂ ਆਦਿ ਵਿਚ ਪਾਇਆ ਜਾਣ ਵਾਲਾ ਜਾਂ ਉਹਨਾਂ ਨਾਲ ਸੰਬੰਧ ਰੱਖਣ ਵਾਲਾ
EXAMPLE	ਸਾਖਰਤਾ ਤੇ ਵਿਚਾਰ-ਵਿਮਸ਼ ਲਈ ਇਕ ਸਮੂਹਿਕ ਸਭਾ ਦਾ ਪ੍ਰਬੰਧ ਕੀਤਾ ਗਿਆ ਹੈ
SYNSET-PUNJABI	ਸਮੂਹਿਕ, ਸਰਵਜਨਕ, ਸਮੁਦਾਇਕ

Figure 4.7: Final result list for the word “आस”

The Algorithm 4.2 is explained using the flow chart. Figure 4.8, shows the flowchart for Hindi to Punjabi dictionary. A concept is represented by same ID number in both Hindi and Punjabi text files and on this basis a word can be searched through mapping of these ID numbers in the text files. First of all, linked list L_1 and L_2 are generated for the text file of Hindi and Punjabi words. When the user enters the word to be searched, the word is searched in the linked list L_1. If the word is found in the list, then the corresponding ID

numbers for each occurrence of the word are retrieved and stored in an array. Again a search operation is performed on the linked list L_2 for each of the ID number stored in array. If the ID match occurs, the corresponding information *i.e.*, the concept, category, example and synonyms of each ID number is fetched and this information is stored in a linked list *F*. Finally this linked list *F* is displayed.

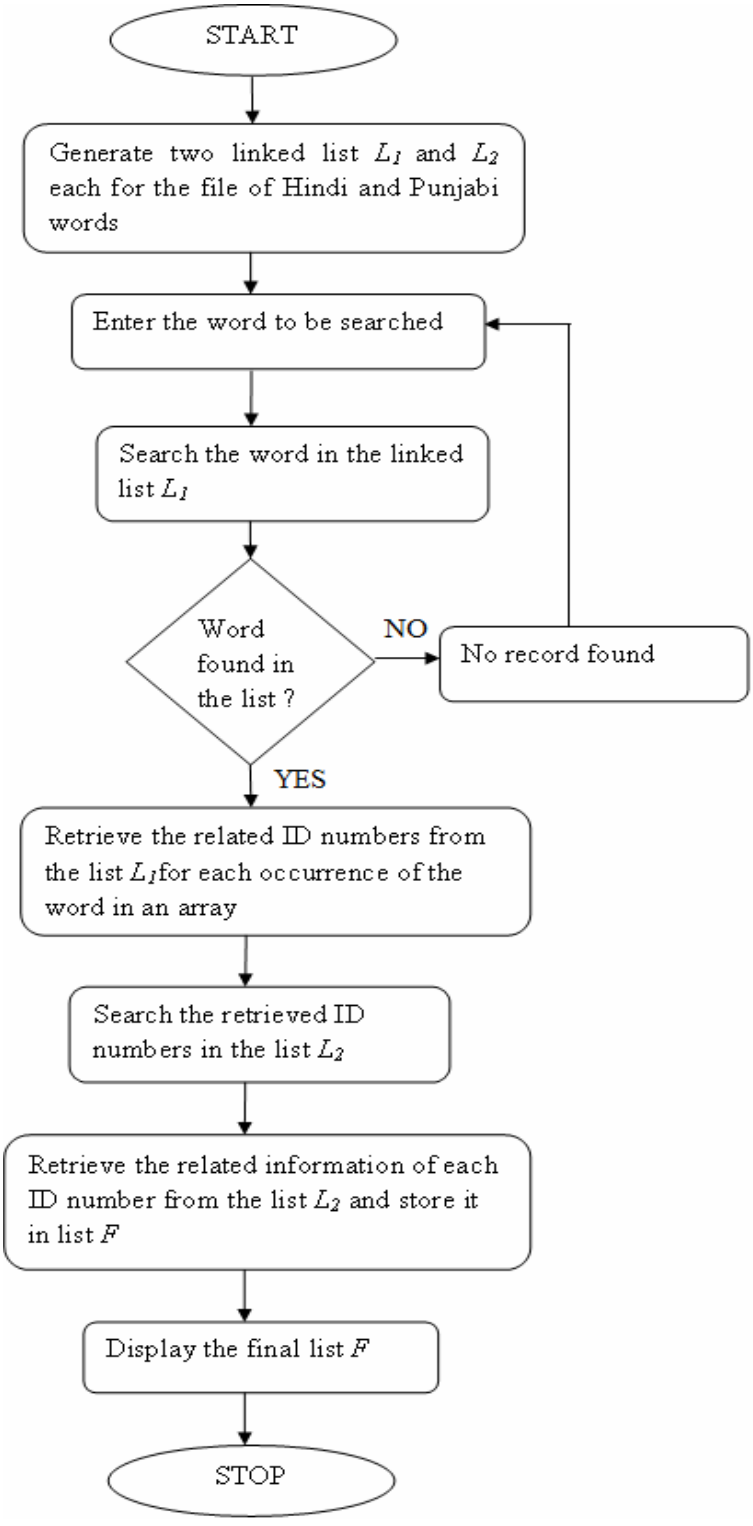


Figure 4.8: Flowchart for Hindi to Punjabi dictionary

4.4 Creation of Punjabi to Hindi Dictionary

Punjabi to Hindi dictionary has around 7168 entries. Each word in the dictionary has an associated information *i.e.,* category like noun, verb etc., concept which explains the word, example usage of the word, and synset and its equivalent in the Hindi language. Punjabi to Hindi dictionary is created using the algorithm as explained in Algorithm 4.2 and using the same methodology as used in developing Hindi to Punjabi dictionary.

`Chapter 5

Experimental Results

The algorithms discussed in Chapter 4 have been used in developing a web application for Punjabi WordNet and Punjabi Hindi bilingual dictionary. The web application developed for Punjabi WordNet is shown in this chapter. In developing this web application, text files obtained as an output of IL-Multidic development tool have been used as the back end and JSP is used as the front end.

5.1 Web Interface for Punjabi WordNet

Figure 5.1 shows the GUI interface of the Punjabi WordNet. Punjabi keypad is used to enter the word which is to be searched. The user can use the View Keypad button to use the Punjabi keypad. Then on the click on the Submit button the various senses of the word in which it can be used will be shown on the page. The number of senses in which it can be used is shown at the top. It will show the category, concept, one example usage and synonymous words for each of the sense. The Next and Previous links can be used to move to the next sense and to the previous sense of the word.

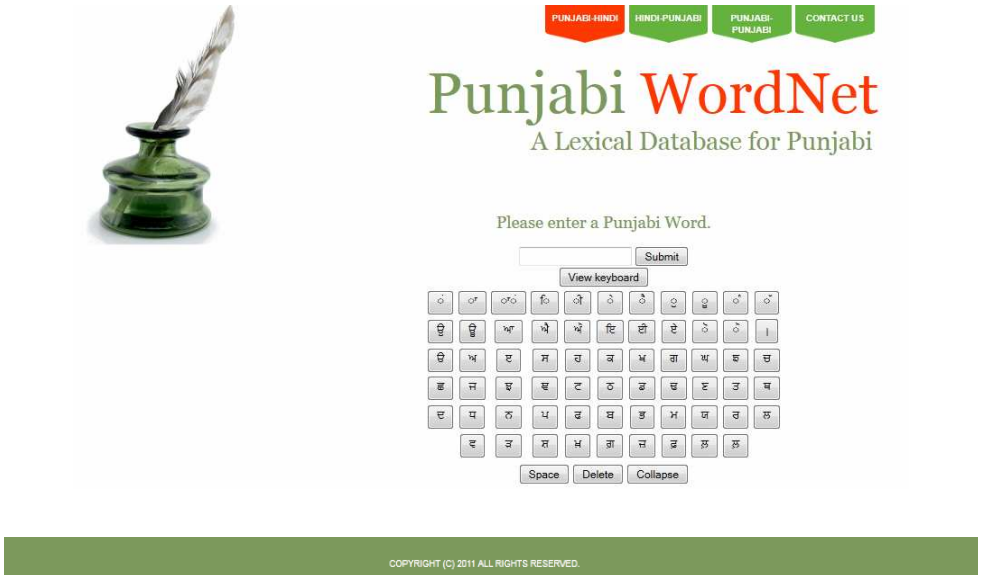


Figure 5.1: Interface for Punjabi WordNet showing Punjabi keypad

For example, if the user types the word "ਪ੍ਰਤੀਕੂਲਤਾ" from the keypad then it will show two different senses in which this word can be used. Figure 5.2 and Figure 5.3 shows two different senses of the word "ਪ੍ਰਤੀਕੂਲਤਾ" in which it can be used.

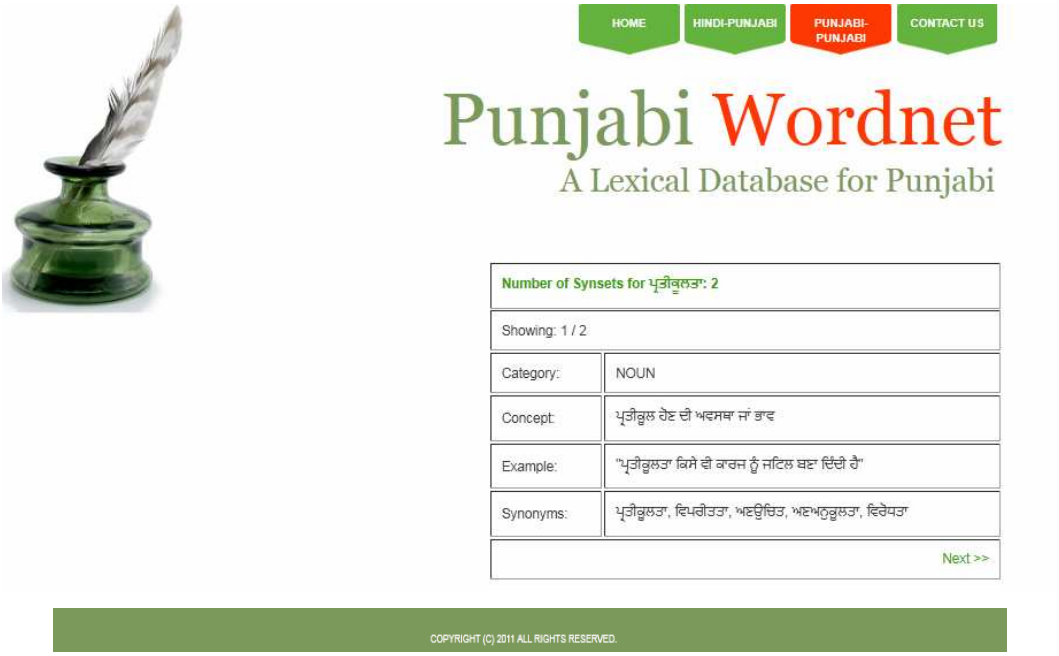


Figure 5.2: Interface showing first sense of the word "ਪ੍ਰਤੀਕੂਲਤਾ"

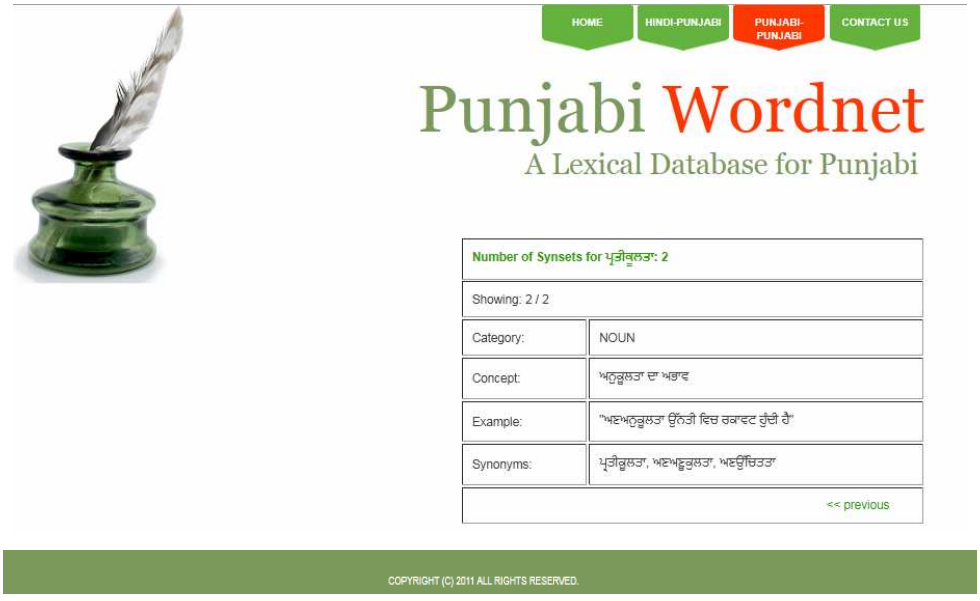


Figure 5.3: Interface showing second sense of the word "ਪ੍ਰਤੀਕੂਲਤਾ"

5.2 Web Interface for Hindi to Punjabi Dictionary

Figure 5.4 shows the GUI interface of the Hindi to Punjabi dictionary. Hindi keypad is used to enter the word which is to be searched. Then on the click on the Submit button the various senses of the word in which it can be used in Punjabi will be shown on the

page. It will show the category, concept, one example usage and synonymous words for each of the sense. The Next and Previous links can be used to move to the next sense and to the previous sense of the word. For example, if the user types the word "आम" from the keypad then it will show four senses in which it is used. Figure 5.5-5.8, shows different senses of the word "आम" in which it can be used in Punjabi.

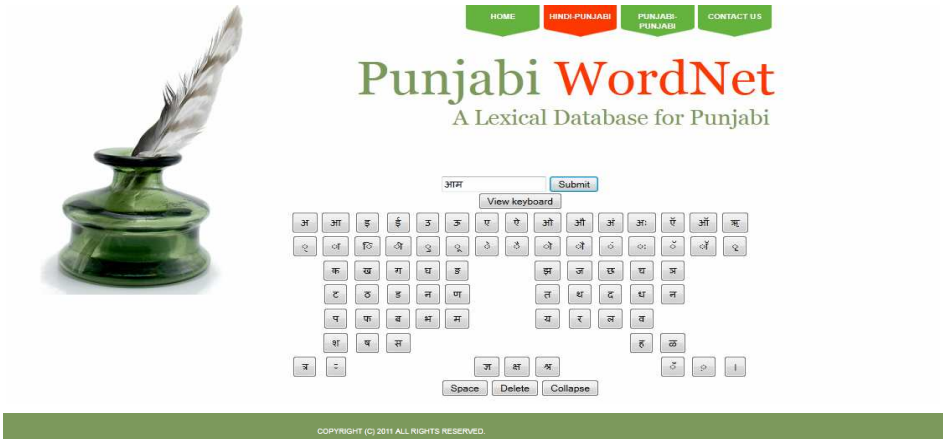


Figure 5.4: Interface showing Hindi keypad for Hindi-Punjabi dictionary

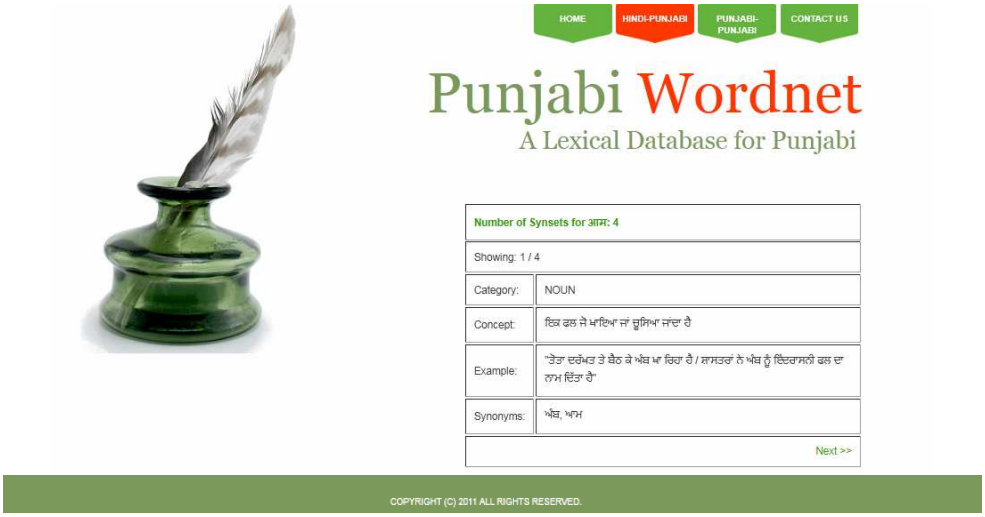


Figure 5.5: Interface showing first sense of the word "आम" in Punjabi

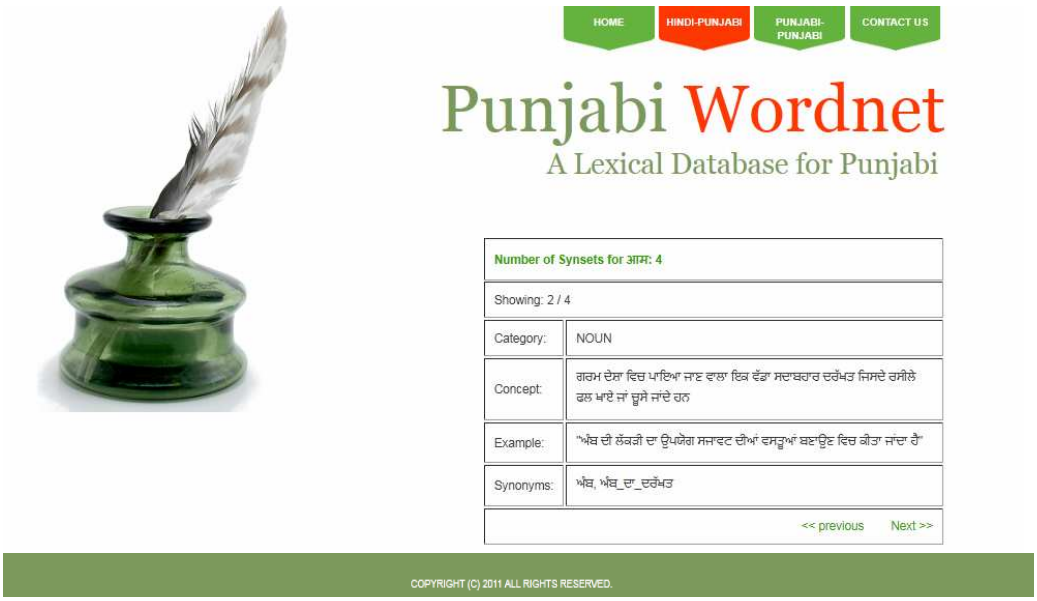


Figure 5.6: Interface showing second sense of the word "आम" in Punjabi

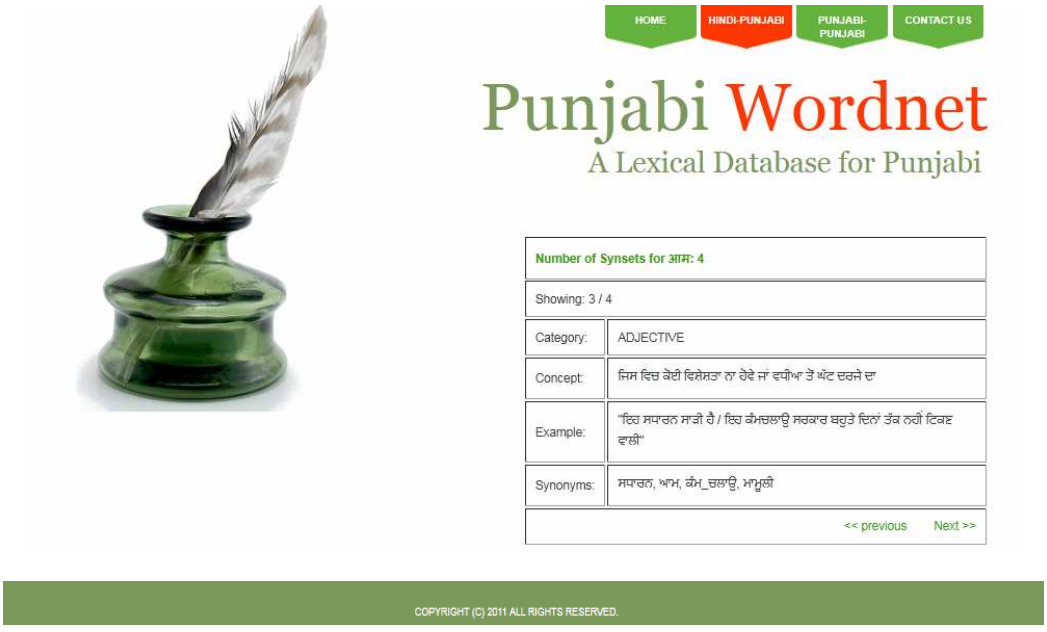


Figure 5.7: Interface showing third sense of the word "आम" in Punjabi

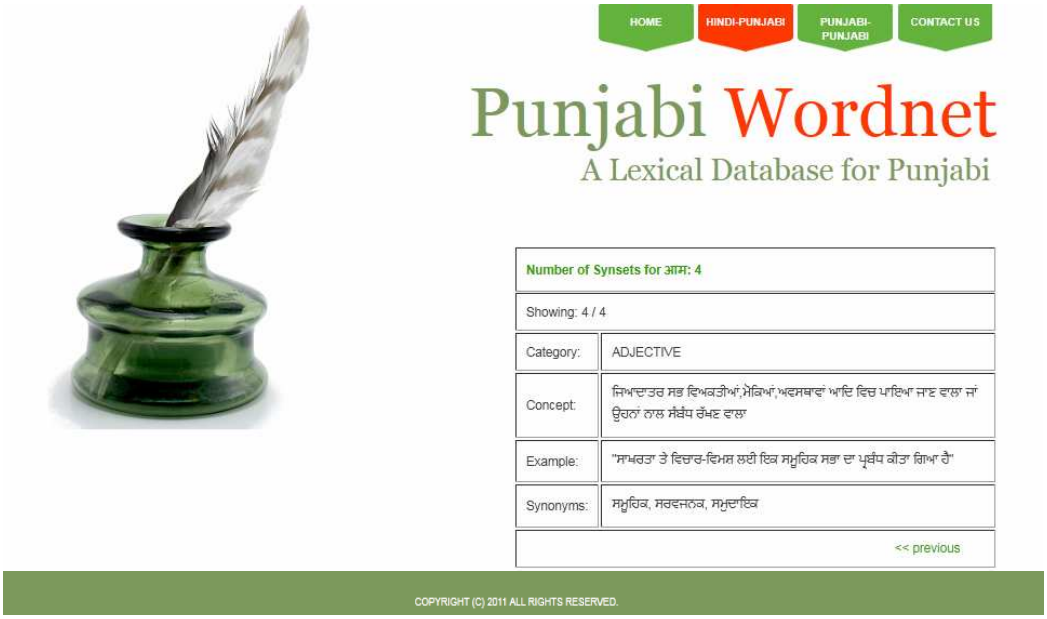


Figure 5.8: Interface showing fourth sense of the word "आम" in Punjabi

5.3 Web Interface for Punjabi to Hindi Dictionary

Figure 5.9 shows the GUI interface of the Punjabi to Hindi dictionary. Punjabi keypad is used to enter the word which is to be searched. Then on the click on the Submit button the various senses of the word in which it can be used will in Hindi be shown on the page. It will show the category, concept, one example usage and synonymous words for each of the sense. The Next and Previous links can be used to move to the next sense and to the previous sense of the word. For example, if the user types the word "ਗਿਆਨ" from the keypad then it will show two senses in which it is used. Figure 5.10-5.14 shows different senses of the word "ਗਿਆਨ" in which it can be used in Hindi.

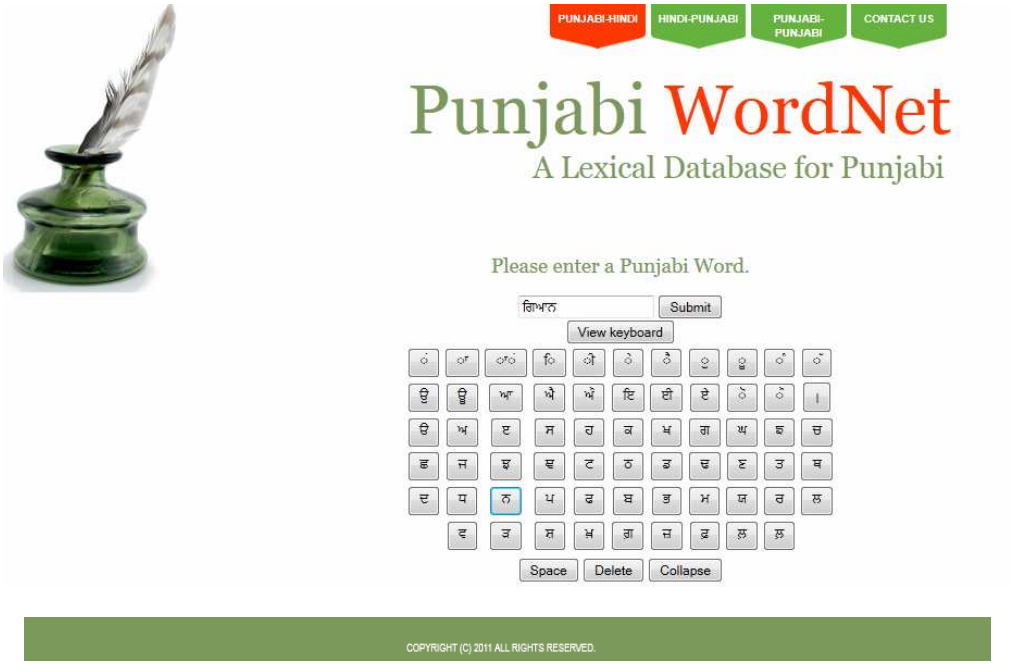


Figure 5.9: Interface for Punjabi to Hindi dictionary

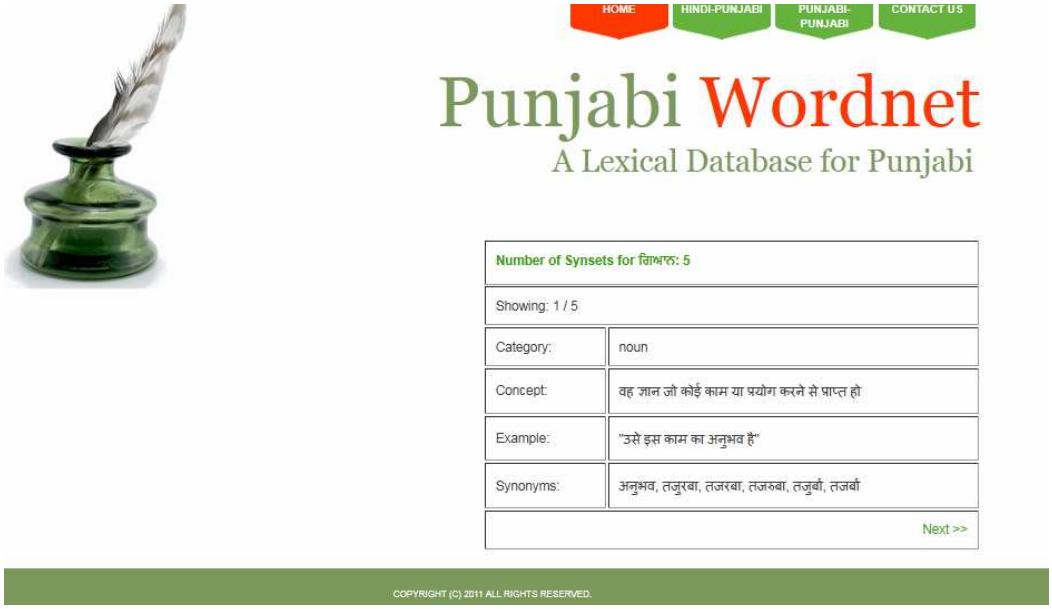


Figure 5.10: Interface showing first sense of the word “ਗਿਆਨ” in Hindi

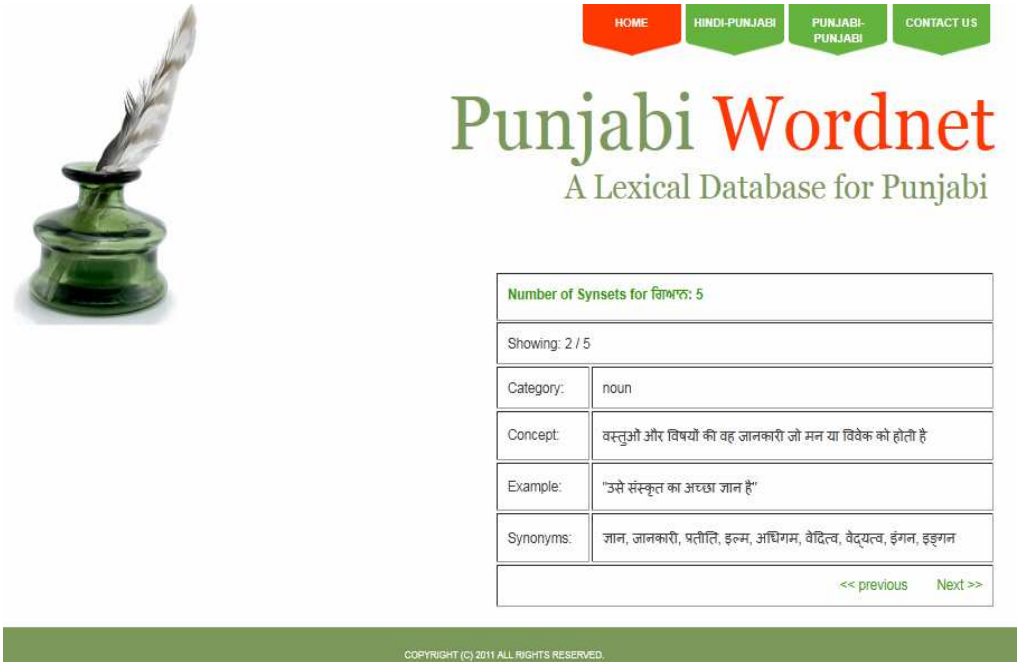


Figure 5.11: Interface showing second sense of the word “ਗਿਆਨ” in Hindi

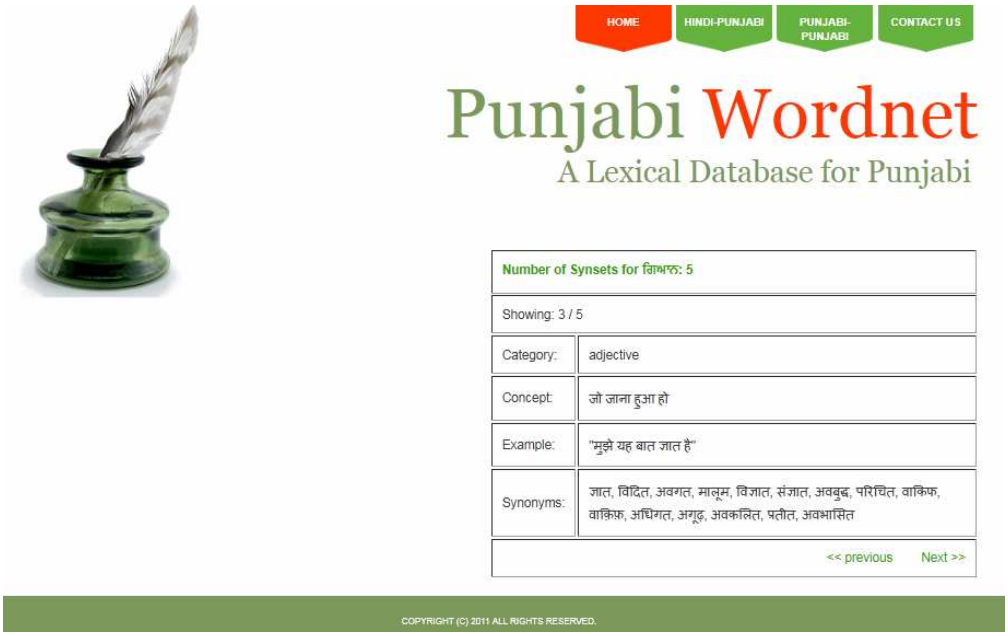


Figure 5.12: Interface showing third sense of the word “ਗਿਆਨ” in Hindi

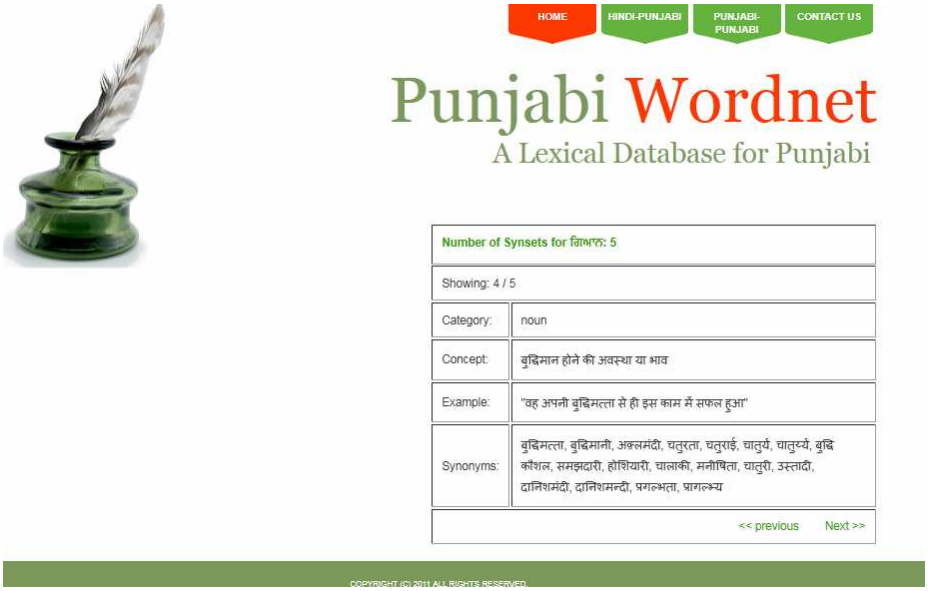


Figure 5.13: Interface showing fourth sense of the word “ਗਿਆਨ” in Hindi

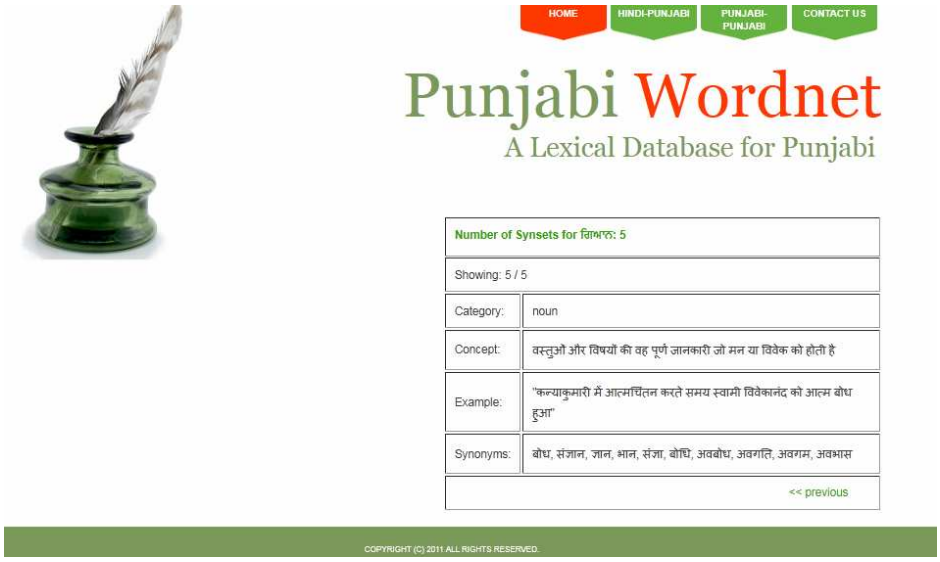


Figure 5.14: Interface showing fifth sense of the word “ਗਿਆਨ” in Hindi

Chapter 6

Conclusion and Future scope

6.1 Conclusion

In this thesis work, an interface of Punjabi WordNet has been developed that can be used as a lexical resource for Natural Language Processing tasks for Punjabi language.

- IL-MultiDic development tool has been used for creation of Punjabi WordNet using Hindi WordNet.
- Punjabi WordNet has been effectively used to develop a Hindi Punjabi bilingual dictionary. Besides providing various senses of a particular word in target language, it will also explain the concept of the word, category of word representing its part of speech, an example sentence and synonymous words.
- A web application for Punjabi WordNet and Hindi Punjabi dictionary has been developed so that these resources can be accessed online, once it is made available on the web.

6.2 Future Scope

There are many possible extensions of this work that can be undertaken in further research. Some of them are given below:

- At present relations like Hypernymy/Hyponymy, Holonymy/Meronymy, Troponymy, Entailment and Antonymy have not been applied on the synsets of Punjabi WordNet. These Lexical and Semantic relations when applied to the synsets will further enrich the present work. These relations can be borrowed from Hindi WordNet as same Id number has been used to represent similar concepts in Hindi and Punjabi WordNet.
- Punjabi WordNet can be used in different Natural language Processing applications *e.g.,* in machine translation, word sense disambiguation, information retrieval *etc*.

- Punjabi WordNet can be used in language teaching and in translation applications.
- The performance can surely be improved if morphology is handled exhaustively. The system currently does not detect the underlying similarity in presence of morphological variations.

References

[1] Brent, M. R., "From grammar to lexicon: Unsupervised learning of lexical Syntax", *Computational Linguistics*, 1993.

[2] Yarowsky, D., Somers, H., Dale, R., Moisl, H., "Word Sense Disambiguation", *Handbook of Natural Language Processing: Techniques andApplications for the Processing of Language as Text*., pp.629-654, 2000.

[3] Aggire, E. and Rigau, G., "Word Sense Disambiguation using Conceptual density",*Proceeding of Coling*, 1996.

[4] Bharati Akshar, Chaitanya Vineet and Sangal Rajeev, "Computational linguistics in India: an overview", *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 2000.

[5] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., "Five papers on WordNet", Princeton University, Cognitive Science Laboratory, Technical report, 1993.

[6] Sinha Manish, Kumar Mahesh, Pande Prabhakar, Kashyap Lakshmi and Bhattacharyya Pushpak, "Hindi word sense disambiguation", *Proceedings of International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India*, 2003.

[7] Rana Shilpa, "Punjabi WordNet –A tool for Natural Language Processing", ME Thesis, Thapar University, Jun. 2010.

[8] Bhattacharya Puspak,"IndoWordNet", *Lexical Resources Engineering Conference, Malta*, 2010.

[9] Hindi WordNet from Center for Indian Language Technology Solutions, IIT Bombay, India. [Online]. Available: http://www.cfilt.iitb.ac.in/WordNet/webhwn

[10] Rana Shilpa, Bhatia Parteek, " PunjabiWordNet-A Tool for Natural Language", *Second National Conference on Recent Advances and Future Trends in IT* ,Feb. 2007.

[11] Kaur Rupinderdeep, Suman Preet, Sharma R.K., Bhatia Parteek, "Punjabi WordNet Relations and Categorization of Synsets", *the 5th Global Wordnet Conference, Mumbai* ,2010.

[12] Punjabi Kosh. [Online]. Available: http://www.punjabikosh.googlepages.com

[13] Punjabi Shabdkosh. [Online]. Available: http://www.4shared.com/file/39293942/9d333376/Punjabi_Shabdkosh__English_to_Punjabi_Dictionary_.html?s=1

[14] Punjabi English dictionary by Punjabi University. [Online]. Available: http://www.advancedcentrepunjabi.org/pedic/Default.aspx

[15] Punjabi English and English Punjabi dictionary. [Online]. Available: http://www.punjabonline.com/servlet/library.dictionary?Action=English

[16] Punjabi encyclopedia and Gurbani dictionary.[Online]. Available: http://www.srigranth.org/servlet/gurbani.dictionary

[17] Shabdkosh [Online]. Available: http://www.shabdkosh.com/

[18] G S Lehal, "A Survey of the State of the Art in Punjabi Language Processing", *Language In India*, vol. 9, no. 10, pp. 9-23, 2009.

Research Publications

- Rekha Rattan, Parteek Bhatia, ”Design and development of Punjabi WordNet and Punjabi Hindi bilingual dictionary”, International Journal of Information and Telecommunication Technology. (Communicated)