AN ONLINE PUNJABI SHAHMUKHI LEXICAL RESOURCE

Ejaz Hasan*, Muhammad Munwar Iqbal, Qaiser Rasool Azeemi*, Ashir Javeed***
* Institute of Computing, Bahauddin Zakariya University, Multan
**Department of Computer Science & Engineering, University of Engineering & Technology, Lahore
**Department of Computer Science, University of Engineering & Technology, Taxila
ejazhasan75@gmail.com, munwariq@gmail.com, qrasool@bzu.edu.pk, ashir.javeed@bzu.edu.pk

ABSTRACT *- Pakistan is a multilingual country where different languages are being spoken by its people. So for the cross lingual information processing there should be some centralized repository of words, usually known as a lexical database. Natural language processing or natural language engineering has many tasks such as word sense disambiguation, machine translation, part of speech tagging and such others. All these tasks also need large scale lexical databases. So there is a rich need to develop such resources. The lexical databases for the developed languages are already built, but less attention is given to less resourced or under resourced languages like Punjabi, Saraiki etc. These are the main motivations behind this research. English WordNet developed by Princeton University is a best example of lexical database. Our work includes design and construction of such a database for Pakistani regional languages. We have studied different approaches adopted for the construction of lexical databases of different languages in the world. In the proposed system, we have developed a web interface that facilitates the updation and query-based results retrieval of entries from lexical database.*

Keywords: Natural Language Processing (NLP), Lexicon or lexical database (LDB). Word Sense Disambiguation (WSD)

INTRODUCTION

Every person is not able to take advantage of heterogeneous information available on the internet as much as it should due to the unawareness from multiple languages or due to the fact that this information is not in the form of user's native language. The new technologies and new theories developed to make all this information accessible more effective as ever before. By using Linguistics and Natural Language Processing (NLP), people made different tools having natural language interfaces to computer systems to make the man-machine interaction easier. The lexical resources got the most importance in computational linguistics and NLP. A very popular lexical resource is English WordNet. A WordNet is a lexical database in which different grammatical categories like nouns, verbs, adjectives and adverbs are grouped together lexically and semantically on the basis of related concepts with identical or closely identical meanings. This research is adopted to make a rich Punjabi lexical database using the similar concepts. Further, the same structure can be adopted for the other local languages.

1.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) is used in two ways by the machines; Natural Language Understanding (NLU) and Natural Language Generation (NLG). In the former, human languages are used to communicate with the computers while in the later Information from the computers are converted in the human languages. In NLP, the text is analyzed by the computer using the available technologies and theories [2]. NLP is possible with the help of some other fields like computer science, psycholinguistics, statistics, electrical and electronic engineering, artificial intelligence, etc. Some of the example applications are machine translation, text summarization, information retrieval etc. [2, 3].

1.1.1 NLP Tasks

When processing natural languages, there are different problems faced by the programmers and technologists. Some of these are discussed below.

- **Text Segmentation**

In the natural text processing system, the major problem faced is text tokenization or text segmentation. How the text is tokenized into words? This can be only possible if we can find the word boundary, which is quite difficult for some languages like Chinese, Japanese and Thai etc.

- **Speech Segmentation**

In the natural or human speech processing system, the major problem faced is sound identification of each character pronounced. Different characters pronounced in different ways on their position in the sentences. There are also problems of word pauses between the successive words in a sentence pronunciation. So to solve such issues, we should design a system which considers the grammar, sentence structure, word senses and the context. Frequency and pitch consideration is also important.

- **Word Sense Disambiguation (WSD)**

Words have different meanings/senses depending upon the particular position in which they are appeared among other words. There should be repository of these words along with their senses. A WordNet or lexical database provides this feature that can be further used in the WSD system.

- **Syntactic Ambiguity**

A natural language sentence is known as syntactic ambiguous in case of having more than one parse trees. The semantic and particular context available for the sentence words can reduce these ambiguities. A lexical resource will help a lot to remove these ambiguities.

- **Machine Translation (MT)**

The translation of human languages amongst others is called machine translation. This is not an easy task but the translation between closely related languages which have some type of similarities like alphabets, structures and grammar, is a bit easy.

- **Named Entities Recognition (NER)**

Named Entities have a significant functionality in NLP Systems. Identification, analysis, extraction, mining and transformation of the named entities like names of persons, organizations, locations, concepts in a given natural language all are challenging NLP tasks [4, 5]. NER is performed through two approaches: linguistic approaches (Rule Based Models) and machine learning approaches e.g. Maximum Entropy Models, Decision Tree, Support Vector Machines, Conditional Random Fields [4]. [5] used Hidden Markov Model (HMM) to perform the NER. Transliteration can also be used for NER [6].

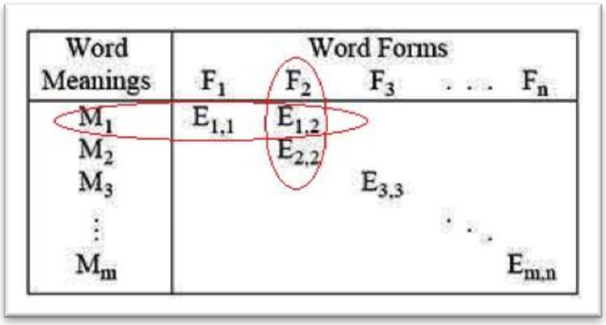
Traditional Databases or Dictionary Databases of Natural Languages are basically the digital versions of the printed dictionary and are based on the primarily the compilation of lexicographer's work and there is no relationships between the words [8]. This is usually known as Machine Readable Dictionary (MRD). Whereas a lexical database (LDB) is a lexical resource focusing on the computational exploitation having a specific structure so that the lexical resource can be used in both NLP systems and human consultation [7, 9].

Literature review

A traditional dictionary usually shows the lexical entries in a sequence alphabetically. Whereas WN is managed on the basis of word meanings; all the words that can represent a particular sense are combined collectively in a synonym set (called synset). All the words of same grammatical category having same concept are grouped together to form the members of that synset. These synsets further connected with each other through the lexico-semantic relations.

2.1.1 Lexical Matrix

The lexical matrix is a vital piece of the human language adopting and learning system. It gives the connection between word form and word meanings. WN structure can be described by the lexical matrix. The word forms referring to the physical utterance denoted the headings for the columns, whereas the word meaning, referring to the lexicalised concept denoted the headings for the rows. For example, as shown in following figure 1, the entry $E_{1,1}$ shows the word form F_1 having the meaning M_1. Also, if there are more than one entries of a particular column, that word form will be polysemous e.g. under the heading F_2 the entries $E_{1,2}$ and $E_{2,2}$ occur so F_2 is polysemous (same word F_2 with multiple senses M_1 and M_2) and if there are more entries in the same horizontal row, the word forms are synonymous (different words with same sense) relative to that context e.g. word forms F1 and F2 are synonyms (entries E1,1 and E1,2 have same sense M1).



Word Meanings	Word Forms				
	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_m					$E_{m,n}$

Figure 1: Elaborating Concept of Lexical Matrix

So some forms might be of polysemous nature (have several different meanings) and several different forms might be of synonymous nature (having same meanings) [11].

2.1.2 Sense and Synsets

Synonymous words are grouped in synsets having same sense or concept. A word has different meanings depending upon the context being used. The context sensitive meaning is called word sense. For example the wordآhas different meanings (or senses) used differently at the different occasions e.g used for (1) calling a person / animal / bird (2) voice at the start of a song / gazal / qawali etc. Also, for a particular word having multiple senses, it will appear in more than one synset pairs. Example the Punjabi word بتّی can have sense وٹّی and بجلی etc.

2.2 Relations in WordNet

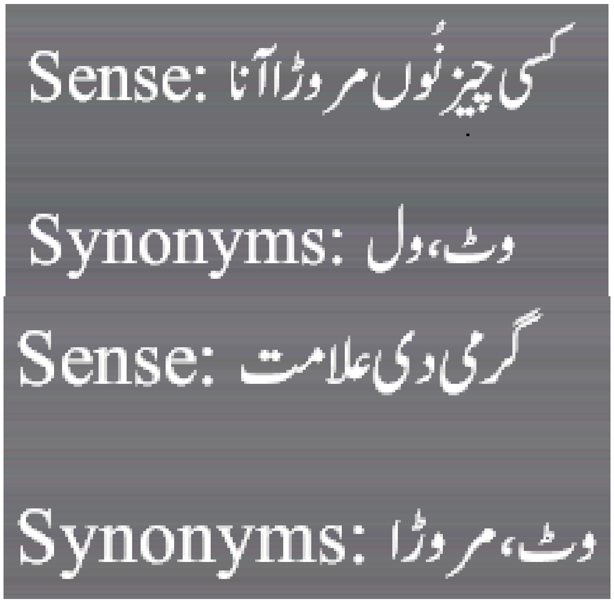
Two types of relations exist: the relations between one synset to another are called semantic relations and the relations between the members of different synsets are called lexical relations.

2.2.1 Semantic Relations

Semantic relations hold between whole synsets. Below we will discuss different semantic relations.

a) Synonymy

WordNets are organized by the concept of synonymy. This is the most important semantic relation. Two words are said to be true synonymous if one word can be used as a substitute for the other in every context. But true synonyms are very hard to find, if they exist anyway. A weakened synonymy definition would be used instead: two words are synonymous in a particular context if one word can be used as substitution for the other in that context and this substitution does not change the overall sentence concept. This relation is symmetric i.e. If one word is similar to other, then other is equally similar to first one. Figure 2 shows some example synonymous groups.

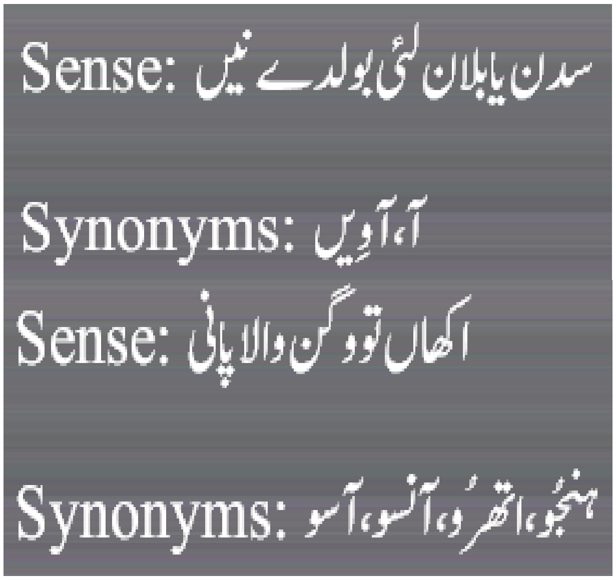


Sense: کسی چیز نوں مروڑ آنا

Synonyms: وٹ،ول

Sense: گرمی دی علامت

Synonyms: وٹ،مروڑا



Sense: سدن یا بلان لئی بولدے نیں

Synonyms: آ، آویں

Sense: اکھاں تو وگن والا پانی

Synonyms: ہنجو، اتھرُو، آنسو، آسو

Figure 2: Example Senses and Synonyms

b) Hypernymy/Hyponymy (a-kind-of or is-a relation)

WordNet is organized in the hierarchy of superset (or superordination) and subset (or subordination) classes. This is a-kind-of relation e.g Red is a (kind of) Color so Red is Hyponymy and Color is its Hypernymy.

c) Meronymy/Holonymy (part-whole or has-a relation)

This relation can be used to maintain a part hierarchy in LDB. Our body has different parts like head, eyes, hands etc. So in this example, body is meronym and body-parts all are its holonyms.

d) Troponymy

A verb does not behave in the same way as noun. Troponymy is a most commonly discovered semantic relation among the verbs of a language. Hyponym ("kind-of" relation) is used for nouns and Troponymy ("manner-of" relation) is used for verbs.

e) Entailment

Entailment is also another semantic relation between verbs of a language. A verb ‘A’ entails a verb ‘B’ if the concept or sense of verb ‘A’ constitutes the concept of the verb ‘B’.

2.2.2 Lexical Relations

Lexical relations hold between members of different synsets. Below we will discuss different lexical relations.

a) Antonymy

Antonymy seems a simple words-connecting relation, but is quite difficult to determine because of its complexity. For example, rich is antonyms of poor, but we cannot say that if someone is not rich it must be then poor. Also usually people say them not rich and not poor. This relation is less occurring among the nouns but is the commonly found relation between the adjectives.

b) Gradation

Gradation is a mechanism to determine the intermediate relation between two antonyms, like noon is between the two antonyms morning and evening.

2.3 Languages of Pakistan

There are many local or regional languages of Pakistan including Punjabi, Pashto, Sindhi, Saraiki, Urdu and Balochi etc. [12, 14]. Most of these languages do not have standardized alphabets. Only Sindhi, Pashto and Urdu have consistent alphabets. Also Urdu or Sindhi alphabets are used to write other languages [14]. In this research, Punjabi is taken as the main consideration. Below we will discuss writing style, grammar and morphology of Punjabi in detail.

Comparison Of Traditional Database and Lexical Database

Table 1: Comparison of Traditional and Lexical Databases

Sr. No.	Attribute	Traditional DB	Lexical DB
1	Computational behaviour	Compilation of Lexicographer’s work	Lexical resource focused on computational exploitation
2	Words Relationships	No relationships between words	Words are grouped together lexically and semantically
3	Usage	Usable by humans Only	Usable both by humans and machines
4	Types	Usually exists in one language	Can be of mono-lingual, bi-lingual and multi-lingual
5	Structure	Sequential alphabetical entries irrespective of word meanings	Arranged with respect to the word meanings
6	Examples	All computer versions of the printed dictionaries	PWN, Kannada WordNet, etc.

2.3.1 Punjabi

The world's twelfth most generally spoken language is Punjabi. Also, it is most common language which is spoken in Pakistan. According Census Report of Pakistan in 2008, number of Punjabi speakers are 76,335,300 which are roughly the 75% of the total population of Pakistan with their first or second language as Punjabi.

a) Punjabi Orthography

Punjabi language has two dialects: Eastern Punjabi mostly spoken by the people of Punjab in India, and Western Punjabi mostly spoken by the people of Punjab in Pakistan [1, 10, 16]. Perso-Arabic (Shahmukhi) script is used by the Pakistani people, and Gurmukhi / Devanagari script is used by the Indian people [12, 13, 15, 17, 21].

Punjabi language connections back with the Indo-Aryan languages [18, 20]. But with the passage of time, Persian, Arabic and Turkish words constitute the Punjabi vocabulary. Also there is a problem of its alphabets. There are no standardized alphabets of Punjabi. It is usually written by using the alphabets of Urdu [14]. Punjabi (especially that is spoken in Pakistan) is a less resourced language. Generally very little work is done on Punjabi [1, 10] (Gurmukhi / Shahmukhi). The main aim behind this research is to create a huge resource of this language in Shahmukhi script that can be used both by the language learners and users and by the linguists to collect certain information for particular NLP applications.

Shahmukhi is written from right to left and is based on Nastalique style of Persian and Arabic script. The shape of the characters in a word is context sensitive, means a letter has different shape if it occurs at the start position, at middle position or end position of a word [17]. This script has thirty eight letters which includes four long vowels Alif (ا) , Vao (و) [v], Choti-ye (ی) [j] and Badi-ye (ے) [j], three short vowels Zer (◌ِ), Pesh (◌ُ) and Zabar (◌َ), diacritical marks likeShad (◌ّ), Khari-Zabar (◌ٰ), do-Zabar (◌ً) [ən] and do-Zer(◌ٍ) [ɪn], or symbol hamza(ء) [13]. Ten aspirated consonants (بھ،پھ،تھ،ٹھ،جھ،چھ،دھ،ڈھ،کھ،گھ) are very frequently used as compared to the remaining six aspirates (رھ، ڑھ، لھ، مھ، نھ، وھ) [22].

The content written in Shahmukhi script generally doesn’t utilize short vowels and diacritical imprints. The same word written with and without diacritics or same word with a different set of diacritics represents the different meanings which make the uncertainty for the machines and for the Punjabi non-speakers especially, e.g. the word ترنا can be written in two ways:تَرنا(To Swim/Float)andتُرنا (To Walk/Move). Similarly the word بیل means ‘the vine’ whereas the word بَیل means ‘the bull’ in English. Thus diacritics are necessary to remove the ambiguities between word meanings/senses [17]. How this was handled in the current implementation? Get the text from the text box, remove the diacritics and search the diacritics free word in the DB, if the word found in DB, get the ID(s) and further search the same word with aerab (diacritics) in the next fields of the searched IDs. OR ask the user with alternate search word with aerabs so that the user can select the exact word otherwise show all the words without aerabs.

b) Punjabi Grammar

Here we will discuss various morphological and syntactic structures of the Punjabi language. Like Urdu, Punjabi also follows the canonical word order of Subject-Object-Verb [16]. There are Masculine and Feminine (two genders), Singular and Plural (two numbers) and six cases, including accusative, nominative, instrumental, ablative, dative, and locative, two types of adjectives (inflected and uninflected), two types of affixes (Prefix and Suffix) [10]. It has postpositions rather than prepositions, e.g. in English we write ‘on the roof’ and in Punjabi it is ‘چھت تے’ or ‘چھت اُتے’.

c) Punjabi-word Classes

Punjabi words have both inflected and uninflected nature. Suffix is mostly used as inflection expressing the grammatical information like number, person, and tense. Nouns are inflected for number and case e.g. مُنڈا ‘boy’ is used for singular and مُنڈے‘boys’ is used for plural. Sometimes the same word used for singular and plural depending upon the context in which it is used e.g.
مُنڈے چھت اُتے چڑھ گئے۔ here مُنڈے word represents the plural category, whereas in مُنڈےنے اپنا گلا کٹ لیا۔, the word مُنڈے represents the singular category.

d) Gender Rule for Punjabi Nouns

If the Noun has ending in alif (ا), it represents a masculine noun, whereas if the ending letter is Choti-ye (ی), the noun will be feminine. For example the word مُنڈا represents the masculine entity and کُڑی represents the feminine entity. Most of the Punjabi words follow this rule with some exceptions.

e) Number Rule for Punjabi Nouns

If the Noun has ending in alif (ا), it represents a singular noun whereas if the word ends at یاں or Badi-ye (ے), the noun will be plural. For example the word مُنڈا represents the singular entity and مُنڈے،کُڑیاںrepresents the plural entities. Most of the Punjabi words follow this rule with some exceptions.

f) Punjabi Adjectives

Adjectives may also be put into inflected and uninflected category. Punjabi adjectives have also inflection for singular and plural, masculine and feminine, etc. We have to contact the appropriate adjective forum that best fits with the noun so following the above mentioned rules for the Noun category, Inflected Adjectives are also marked through endings, for the gender, for the number and the noun cases they qualify. For Example کالا بکرا, word کالا is masculine adjective and word بکرا is a masculine noun;بکریکالی, word کالی is feminine adjective and بکری is a feminine noun, which are in accordance with the gender of the noun it qualifies in both cases. The uninflected adjectives are totally constant and rigid having ending in either consonants or vowels, for Example بدشکل ،اُداس.

g) Tonal Features of Punjabi

Modern Punjabi has a phonetic nature and due to this nature it is called a tonal language [16], means the same word with different pronunciation make a different word/sense e.g. the word کوڑا can be pronounced as KoRa (horse), KoRRa (leprosy/a disease), KoRaa (whip) . The tonal / melodic features are the inherent properties of pronunciation of a word. There are three tones in Punjabi: high-falling / high tone, low-rising / low tone and mid tone / level tone [19].

Problem Statement

Translation from one language to another is not an easy task due to the ambiguous nature of natural languages. The symbol system is quite easy for humans to learn but is very difficult for computers because of complexity and different usage of the human language strings. These strings or sentences are composed of words that contain the language concepts and other lexical and semantic information associated with them [23]. This is a little effort towards suggesting feature incorporation suited to lexical database of Pakistani local languages for NLP applications. The main aim of the thesis is to build a Punjabi Lexical Database and providing a web interface that can be accessed online.

RESEARCH METHODOLOGY

We are using MySQL database tables at the backend, to store the lexical and semantic information of our local and regional languages. The frontend has two views, one for the DB manager, who can store and update DB; the other view for the end-user, who can query the word. We have stored words, their forms with possible part of speech tags, senses, set of synonyms (synsets) based on the word sense, gloss and example sentences showing the typical usage of an entry in LDB. In addition, the word forms are linked with the main word along with their own unique and distinguishable sense. For example in Punjabi, the word وٹ have many forms like متھے تے وٹ، کپڑے تے وٹ، ڈھڈ وچ وٹ showing different senses for each pair. We are not going to store the phonetic and phonemic features usually required by the speech processing system, at this time. There is no morphological model embedded in the construction of current lexical database.

This structure is proposed and implemented to keep the aim basically to create an online repository for language users and language learners and in future for creating dictionaries and Interlingua translation systems based on this online resource.

The overall research is divided into different phases or steps. Before starting work in these phases, the existing LDB and WordNets are studied and their structures and functionalities are carefully noted for the implementation of our LDB.

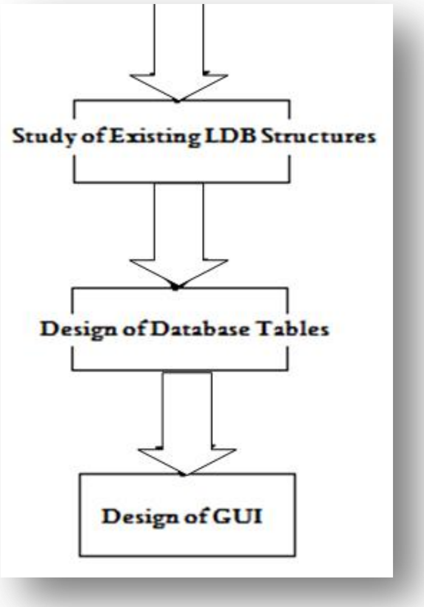


Figure 3: Methodology Adopted (Step 1)

In the first phase a database and a GUI was designed. This database was not a finalized version, as some modification needed will be added in the preliminary design. At the start, 1000 Punjabi words and their related forms are stored with grammatical categories and senses. In the next phase all the words and their senses are carefully read and the synsets were made. In the third phase different other semantic and lexical relations were be implemented and checked via queries to the lexical database. These phases were followed again and again for the next entries.

Proposed Solution:

Web interface developed for this project is shown below. A text box is given where a user can write a Punjabi word and click 'Search' button.

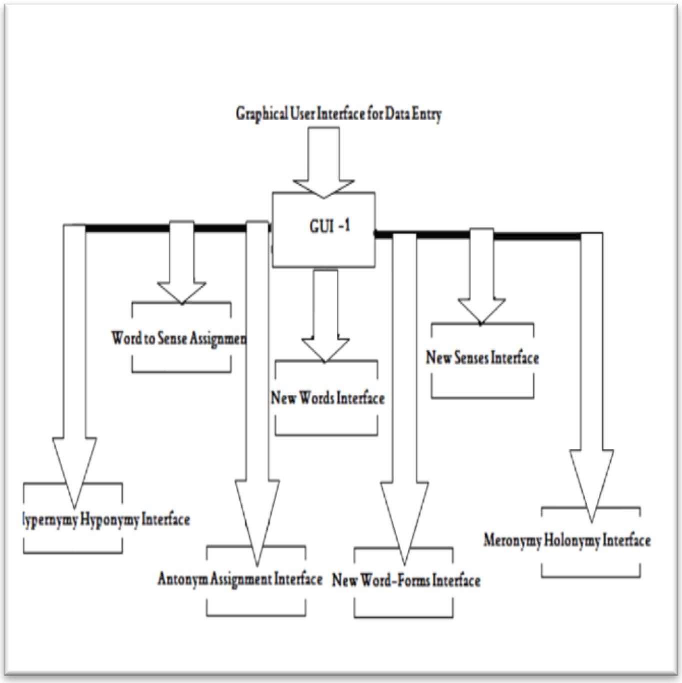
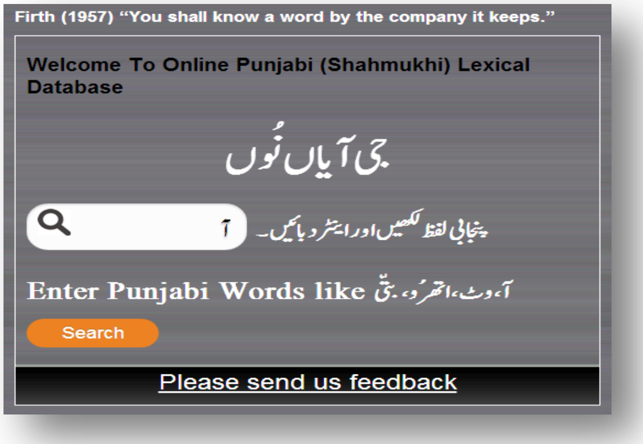

Graphical User Interface for Data Entry
GUI -1
Word to Sense Assignmen
New Words Interface
New Senses Interface
lypernymy Hyponymy Interface
Antonym Assignment Interface
New Word-Forms Interface
Meronymy Holonymy Interface

Figure 4: Methodology Adopted (Step 2)


Firth (1957) "You shall know a word by the company it keeps."

Welcome To Online Punjabi (Shahmukhi) Lexical Database

جی آیاں نُوں

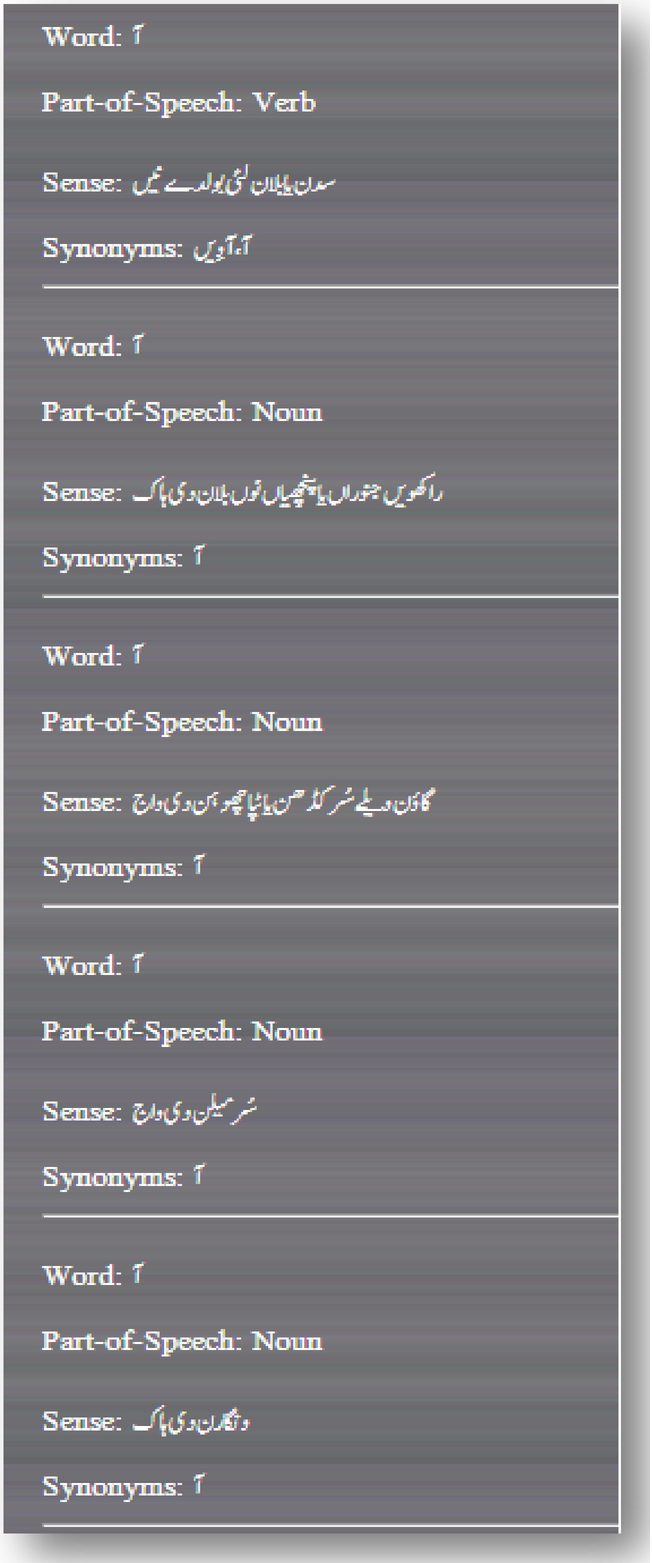
آ پنجابی لفظ لکھیں اور اینٹر دبائیں۔

Enter Punjabi Words like آ،وٹ،اتھرُو،بتّی

Search

Please send us feedback

Figure 5: Front-end-User Interface



Word: آ

Part-of-Speech: Verb

Sense: سدان پایلان لئی بولدے نیں

Synonyms: آ،آویں

Word: آ

Part-of-Speech: Noun

Sense: راکھویں جنوراں یا پنچھیاں نوں بلان دی ہاک

Synonyms: آ

Word: آ

Part-of-Speech: Noun

Sense: گاؤن ویلے سُر کلا حسن یا پچھوہن دی واج

Synonyms: آ

Word: آ

Part-of-Speech: Noun

Sense: سُر میلن دی واج

Synonyms: آ

Word: آ

Part-of-Speech: Noun

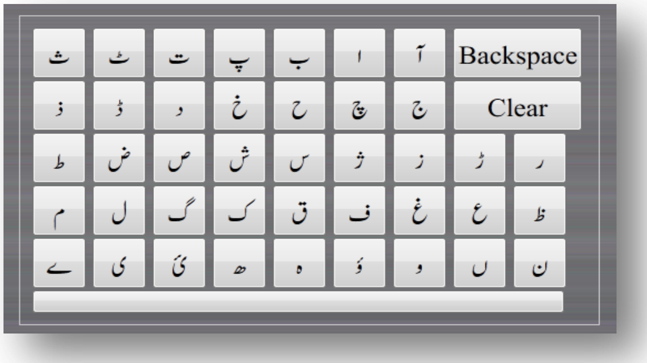
Sense: ونگاران دی ہاک

Synonyms: آ

Figure 6: Frontend-Showing Senses and Synonyms of word (آ)

The word is searched in the database word-table and retrieved its id. The word-id is then searched in the assign-word-sense-table and retrieved all the respective sense-ids. These sense-ids represent the unique senses in the sense-table from where the sense-text is retrieved. In this way, the given input word, all its senses and words corresponding to these senses are retrieved. For example, when the word آ is searched the following output will be shown.

A user can enter the Punjabi words either by using physical keyboard or on screen keyboard (OSK). When a Punjabi word is being searched, the result set consisted of word, grammatical category, sense and set of synonym words will be shown.



ث	ٹ	ت	پ	ب	ا	آ	Backspace	
ذ	ڈ	د	خ	ح	چ	ج	Clear	
ط	ض	ص	ش	س	ژ	ز	ڑ	ر
م	ل	گ	ک	ق	ف	غ	ع	ظ
ے	ی	ئ	ھ	ہ	ؤ	و	ں	ن

Figure 7: On Screen Keyboard

4.1 CONCLUSION AND FUTURE SCOPE

In this thesis a web interface was developed for the Punjabi Lexical Database where a user can enter a Punjabi word for searching different attributes of this word. The developed system provides grammatical category, word senses, synonyms and example sentences showing the brief concept of the synsets. This LDB can also be used for various NLP applications. All the entries were made by hand; there is no automatic creation or extraction of lexical database entries. All the word related information like part-of-speech, the concept underlying, example sentences, glosses etc. is provided for each lexical entry. There is no restriction to enter all this information at once. The language expert can enter individual information related to a particular entry at various time. India has already developed Punjabi grammatical error detection system, Punjabi WordNet and Punjabi-Hindi bilingual dictionary. But those systems are not available openly, that is why we have to develop our own LDB for Punjabi and other local languages. Also, those systems are not using the Shahmukhi scripts rather they are entirely using the Gurmkhi or Devanagri scripts. We have developed and saved the system data in Shahmukhi scripts using standard Unicode. There are numerous potential applications of this research that can be attempted in further research titles.

REFERENCES

1. Narang, Ashish, R. K. Sharma, and Parteek Kumar. "Development of Punjabi WordNet." CSI transactions on ICT 1.4 (2013): 349-354.
2. Liddy, Elizabeth D. "Natural language processing." (2001).

3. Chowdhury, Gobinda G. "Natural language processing." Annual review of information science and technology 37.1 (2003): 51-89.
4. Gupta, Vishal, and Gurpreet Singh Lehal. "Named Entity Recognition for Punjabi Language Text Summarization." International Journal of Computer Applications 33.3 (2011): 28-32.
5. Morwal, Sudha, and Deepti Chopra. "NERHMM: A Tool For Named Entity Recognition based on Hidden Markov Model." International Journal on Natural Language Computing (IJNLC) Vol 2: 43-49.
6. Morwal, Sudha, Deepti Chopra, and G. N. Purohit. "Named Entity Recognition in Natural Languages Using Transliteration."
7. JANSSEN, MAARTEN. "Lexical vs. Dictionary Databases."
8. Hong, Jer Lang. "MalayWordNet—A novel lexical database for Malay language." Fuzzy Systems and Knowledge Discovery (FSKD), 2013 10th International Conference on. IEEE, 2013.
9. Sáenz, F., and A. Vaquero. "Applying relational database development methodologies to the design of lexical databases." Database Systems (2005): 231-238.
10. Kaur, R., Sharma, R. K., Preet, S., & Bhatia, P. (2010). Punjabi WordNet relations and categorization of synsets. In 3rd national workshop on IndoWordNet under the aegis of the 8th international conference on natural language processing (ICON 2010), Kharagpur, India.
11. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. International journal of lexicography, 3(4), 235-244.
12. http://lrwiki.ldc.upenn.edu/mediawiki/index.php/Panjabi/Panjabi; visited on 12-08-2014
13. Saini, Tejinder Singh, and Gurpreet Singh Lehal. "Word Disambiguation in Shahmukhi to Gurmukhi Transliteration." Asian Language Resources collocated with IJCNLP 2011 (2011): 79.
14. Bhurgri, Abdul-Majid. "Enabling Pakistani Languages through Unicode." Microsoft Corporation white paper at http://download. microsoft. com/download/1/4/2/142aef9f-1a74-4a24-b1f4-782d48d41a6d/PakLang.pdf (2006).
15. Virk, Shafqat Mumtaz, Muhammad Humayoun, and Aarne Ranta. "An Open Source Punjabi Resource Grammar." RANLP. 2011.
16. Sharma, Dharam Veer. "An Analysis of Difficulties in Punjabi Language Automation due to Non-standardization of Fonts." (2011).
17. Malik, Muhammad G. "Punjabi machine transliteration." Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006.
18. Gill, Mandeep Singh, and Gurpreet Singh Lehal. "A grammar checking system for punjabi." 22nd International Conference on on Computational Linguistics: Demonstration Papers. Association for Computational Linguistics, 2008.
19. Baart, Joan LG. "Tonal features in languages of northern Pakistan." Pakistani languages and society: problems and prospects (2003): 132-144.
20. Dhillon, Rajdip. "STRESS IN PUNJABI."
21. Dua, Mohit, et al. "Punjabi automatic speech recognition using HTK." IJCSI International Journal of Computer Science Issues 9.4 (2012): 1694-0814.
22. Malik, Muhammad G. "Punjabi machine transliteration." Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006.
23. T.H. King, S. Dipper, A. Frank, J. Kuhn, and J. Maxwell, “Ambiguity Management in Grammar Writing”, 2000