ISSN:0975-9646
Dharam Veer Sharma et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (4) , 2011, 1434-1437

Punjabi Language Characteristics and Role of Thesaurus in Natural Language processing

Dharam Veer Sharma[1] Aarti[2]
Department of Computer Science, Punjabi University, Patiala, INDIA

***Abstract*---This paper describes an attempt to explain various characteristics of Punjabi language. The origin and symbols of Punjabi language are presents in this paper. Various relations exist in thesaurus and role of thesaurus in natural language processing also has been elaborated in this paper.**

***Keywords*---Thesaurus, Punjabi, characteristics, relations**

1. INTRODUCTION

A thesaurus links semantically related words and helps in the selection of most appropriate words for given contexts [1]. A thesaurus contains synonyms (words which have basically the same meaning) and as such is an important tool for many applications in NLP too. The purpose is twofold: For writers, it is a tool - one with words grouped and classified to help select the best word to convey a specific nuance of meaning, and to provide vocabulary (or terminology control) when there are several possible terms designating a single concept, the thesaurus should lead the writer, indexer and searcher to the appropriate concept, regardless of the word with which they started initially.

Punjabi is the language used by hundreds of millions of people in India, and is also the language used by Punjabis around the world [2]. Surprisingly, little has been done in the field of computerization and lexical resources of this language [4]. It is therefore motivating to develop a Punjabi language thesaurus.

This paper is divided into six sections. Section 2 gives a brief account on the morphological, syntactical and technical characteristics of Punjabi language. Section 3 provides role of thesaurus in Natural Language Processing (NLP). Section 4 discusses various relationship exist in thesaurus (for document processing). Section 5 discusses various phases of Punjabi thesaurus. Section 6 concludes the work presented in this paper.

2. PUNJABI LANGUAGE

2.1 Origin and Symbols

Punjabi is classified as a member of the Indo-Aryan subgroup of the Indo-European family of languages. The Punjabi language is a descendent of the Sauraseni Prakrit, a language of medieval northern India. It believed to have developed as a distinct language from the Shauraseni Apabhramsha language around the 11th century. Other early influences on Punjabi include Indo-Aryan and pre-Indo-Aryan languages [5].

India is a country of 122 languages; among these 22 are official languages declared by Government of India [6]. Punjabi language is world's 12th most widely spoken language. Punjabi Language is used in both parts of Punjab, in India and also in Pakistan. Punjabi is syllabic in nature. It consists of 41 consonants called *vianjans*, 9 vowel symbols called *laga* or *matras* and 2 symbols for nasal sounds (. , °) [7][8].

2.2 Characteristics of the Punjabi Language

Modern Punjabi is a very tonal language, making use of various tones to differentiate words that would otherwise be identical. Three primary tones can be identified: high-rising-falling, mid-rising-falling, and low rising. Following are characteristics of Punjabi language [3] [4].

2.2.1 Morphological characteristics

Morphologically, Punjabi is an agglutinative language. That is to say, grammatical information is encoded by way of affixation (largely suffixation), rather than via independent freestanding morphemes. Punjabi nouns inflect for number (singular, plural), gender (masculine, feminine), and declension class (absolute, oblique). The absolute form of a noun is its default or uninflected form. This form is used as the object of the verb, typically when inanimate, as well as in measure or temporal (point of time) constructions. There are seven oblique forms in Punjabi, corresponding more or less to the case forms: agentive/nominative, accusative-dative, instrumental, ablative, genitive, locative, and vocative. All cases except for the vocative are distinguished by means of postpositions. The vocative form takes no postposition, but may be preceded by a vocative particle or term of address. Punjabi verbs inflect for tense, aspect (perfective, imperfective), mood (indicative, imperative, and subjunctive, conditional), voice (active, passive), person, number, and gender. In this way, Punjabi verbs agree with their subjects, as is the case with other Indic languages. Adjectives inflect for gender and number and thus agree with the nouns they modify. Adverbs do not inflect. With respect to morphology, Punjabi and Gujarati are nearly identical.

2.2.2 Syntactic Characteristics

General syntactic structure of Punjabi language is Subject, Object and Verb (SOV). Punjabi sentences are mainly simple in structure but complex and compound sentences are also found in literature. Punjabi sentence structure is flexible. Depending on the context or mood of the speaker, it might vary. Punjabi sentences are mostly analytic in structure but the feature of synthesis is still found at dialectal level.

2.3 Technical Characteristics

Each Language has its own script suited to its particular needs and there are certain rules governing its writing and usage. These rules pertain to its script, vowel-signs, correct pronunciation, numerals, spellings and its dialectical variations, punctuation marks, phonemes and their nearest equivalents in other languages, standard terminology, forms of verbs, declensions and other grammatical subtleties.

2.3.1 Punctuation Marks

Punctuation marks are symbols that indicate the structure and organization of written language, as well as intonation and pauses to be observed when reading aloud. By analyzing a syntactic construction and word formation, punctuation marks delimiting them to the concerned fractions explicate the underlying meaning perfectly. In the ancient Punjabi

single full stop bar (.) or double full stop bar (..) had been in vogue only. However, to the present day Punjabi besides Dandi (।) many other marks are used out of which the mark of Dandi (।) has been retained from old Punjabi while all other marks have been derived from English.

2.3.1.***1 Full Stop***

A full stop or period is the punctuation mark commonly placed at the end of sentences. There are four types of marks used for full stop in Punjabi i.e. Dandi (।), double Dandi (॥), Sign of interrogation (?) Sign of exclamation (!) They primarily denote full stop.

A) Dandi (।)

In place of full stop of English, Dandi is used in Punjabi. It is put after the descriptive and imperative sentences.

ਮੇਰਾ ਨਾਮ ਆਰਤੀ ਹੈ ।

B) Double Dandi (॥)

In the ancient Punjabi, the use of double Dandi was customary at the end of the sentences but in present Punjabi writing system Dandi (।) is used only.

C) Sign of Interrogation (?)

The question mark (?; also known as an interrogation point, interrogation mark, question point), is a punctuation mark that replaces the full stop (period) at the end of an interrogative sentence in Punjabi and many other languages. The question mark is not used for indirect questions. The question mark character is also often used in place of missing or unknown data. For example:

ਤੂੰ ਕਿੱਥੇ ਕੰਮ ਕਰਦਾ ਹੈਂ ?

However in the indirect interrogative sentences the Dandi (।) will be put instead of interrogation mark such as-

ਉਸ ਨੇ ਮੈਨੂੰ ਪੁਛਿੱਆ ਕਿ ਮੈਂ ਕਿੱਥੇ ਕੰਮ ਕਰਦਾ ਹਾਂ ।

D) Sign of Exclamation (!)

An exclamation mark, exclamation point, or bang (!) is a punctuation mark usually used after an interjection or exclamation to indicate strong feelings or high volume (shouting), and often marks the end of a sentence. For example:

ਵਾਹ! ਕਿੰਨਾ ਸੁੰਦਰ ਨਜ਼ਾਰਾ ਹੈ ।

2.3.1.2 ***Comma (,)***

The **comma** (,) is a punctuation mark. It has the same shape as an apostrophe or single closing quotation mark in many typefaces, but it differs from them in being placed on the baseline of the text. In general, the comma is used where ambiguity might otherwise arise, to indicate an interpretation of the text such that the words immediately before and after the comma are less closely or exclusively linked in the associated grammatical structure that they might be otherwise. For example

ਤੂੰ, ਮੈਂ ਅਤੇਂ ਉਹ

2.3.1.3 ***Semi-Colon (;)***

The pause in this punctuation mark is more than that of comma and half than that of full stop. It is used in the compound sentence to distinguish the clause with commas or in the long sentences in order to distinguish the least connected clauses. Semicolons indicate a stronger separation than a comma but weaker than a period. For example:

ਹੌਲੀ ਗੱਲਾਂ ਕਰੋ ; ਕੰਧਾਂ ਦੇ ਵੀ ਕੰਨ ਹੁੰਦੇ ਹਨ |

2.3.1.4 ***Colon (:)***

The duration of pause is longer in colon then that of semicolon. It is used in a sentence to differentiate the quantitative details or illustration section or it is used in the independent sentences having same drift.

ਸੌਣੀ ਦੀਆਂ ਮੁੱਖ ਫਸਲਾਂ ਹਨ : ਝੋਨਾ, ਕਪਾਹ, ਮੱਕੀ ਅਤੇ ਦਾਲਾਂ

2.3.1.5 ***Colon dash (:-)***

This sign is used before giving any quotation or the following details or any instance or after giving the heading of the paragraph in the same line.

ਹੇਠਾਂ ਲਿਖੇ ਅਨੁਸਾਰ :-

2.3.1.6 ***Dash (-)***

A dash is one of several kinds of punctuation mark. Dashes appear similar to hyphens, but vary to them in appearance and serve different functions. The most common versions of the dash are the en dash (–) and the em dash (—).

ਤੁਹਾਡਾ ਬਹੁਤ-ਬਹੁਤ ਧੰਨਵਾਦ |

2.3.1.7 ***Inverted Commas (“….”)***

These Commas are used in the beginning and end of an utterance or sentence that is given exactly.

2.3.1.8 ***Apostrophe (‘)***

Apostrophe is used against that part of the phoneme which is left out at the time of articulation.

‘ਵਾਜ਼ (ਅਵਾਜ਼)

2.3.1.9 ***Bindi (.)***

Like English full stop, Bindi in Punjabi is used to put any word in an abbreviated form.

ਮੈਂ ਐਮ.ਏ. ਪਹਿਲੇ ਸਾਲ ਵਿੱਚ ਹਾਂ ।

2.3.1.10 ***Jorni (Hyphen) (-)***

It is used in the copulative compound.

1993-2010, ਇੱਕ - ਦੱਸ

2.3.1.11 ***Oblique line (/)***

An oblique line is drawn between the two alternative words. For example:

ਸ਼੍ਰੀ /ਸ਼੍ਰੀਮਤੀ

ਪੁੱਤਰ/ਪੁੱਤਰੀ

2.3.1.12 ***Brackets or Parentheses ( ), { }, [ ]***

The word meaning, other forms, some specific information or the additional part of the syntax is given in the parentheses; such as

100/ ਰੁ (ਇੱਕ ਸੌ ਰੁਪਏ)

3. ROLE OF THESAURUS IN NATURAL LANGUAGE PROCESSING

Thesaurus is considered to be the most important resource available to writers of documents, text analysis, and in many related areas. Natural language processing is essential for dealing efficiently with the large quantities of text [5]. The most common use of natural generation technology is to create computer system that present information to people in a representation that they find easy to comprehend. All types of NLP (Natural Language Processing) tasks need a thesaurus. When user works on documents by using natural

language like Punjabi, Hindi, Oriya, Bengali etc, system needs to pick best word to convey a specific nuance of meaning. Writer without using a thesaurus means that different descriptive terms can be assigned to resources about the same subject. The importance of thesaurus in natural language will be clearer from following example. Consider:

- ਉਸਤਤ
- ਉਪਮਾ
- ਪ੍ਰਸ਼ੰਸਾ
- ਸ਼ਲਾਘਾ
- ਵਡਿਆਈ

All refer to the concept of word 'praise'. If no control is placed on subject keywords, any of these might be used by an indexer to describe a resource about praise. Similarly, users searching for information or resources commonly define the same query using differing terms. If there is no guidance to one term from a set of synonymous terms through the use of a thesaurus, users may not be able to locate all the resources that are relevant to their search.

4. RELATIONS USED IN THESAURUS

Thesaurus groups words which have same meaning in one category. In similar manner, also groups set of antonyms under one category. The central objective of thesaurus is better document writing so that word presents most meaningful and correct syntax of a sentence.

4.1 Synonyms

Synonymy means similarity of meaning. For thesaurus building, the task is to calculate similarity between words on the basis of the grammatical relations they both share. 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*. Following words represent the synonym relation between the words [3].

- For example the word ਉਦਾਸ (sad) has synonyms ਚਿੰਤਾਤੁਰ, ਉਪਰਾਮ, ਨਿਰਾਸ਼, ਪਰੇਸ਼ਾਨ, ਨਾਖੁਸ਼ etc.
- Similarly following are few words with synonyms. ਉਤੇਜਕ (excited), ਭੜਕਾਊ, ਭੱੜਕਾਵਾਂ, ਉਕਸਾਉ
- ਉਸਤਤ (praise) , ਉਪਮਾ, ਪ੍ਰਸੰਸ਼ਾ, ਸ਼ਲਾਘਾ etc.

4.1.1 Words which have synonyms with more than one context

There are some words which have synonyms under more than one context. Following are some words example of current category.

- ਖ਼ਰਾਬ(dirty) has synonyms ਗੰਦਾ, ਮੈਲਾ, ਮਿਟਿਆਲਾ, ਮਲੀਨ etc. The word ਖ਼ਰਾਬ has another context also. The meaning of ਖ਼ਰਾਬ is dirty which come under one context, and another meaning of ਖ਼ਰਾਬ (mean) has synonyms ਬੁਰਾ, ਪਤਿਤ, and ਘਟੀਆ under another category.
- The word ਝੂਠਾ (liar) has synonyms ਫਰੇਬੀ, ਮੱਕਾਰ and ਛਲੀ under one context and ਝੂਠਾ (showoff) has synonyms ਨਕਲੀ, ਦਿਖਾਵਟੀ and ਜਾਅਲੀ under another context.

4.1.2 Words which can be write in more than one way (Homonyms words)

There is no standardization of Punjabi spellings. A word may be spelled in more than one way and all the forms may be acceptable.

- ਆਪਣਾ, ਆਪਨਾ
- ਬਿਪਰੀਤ, ਵਿਪਰੀਤ
- ਗੂੜਾ, ਗੂੜ੍ਹਾ
- ਹਨੇਰਾ, ਹਨ੍ਹੇਰਾ
- ਅਨੇਖਾ, ਅਨੈਖਾ
- ਆਰੰਭ, ਅਰੰਭ
- ਵਰਕ, ਬਰਕ
- ਜਤਨ,ਯਤਨ etc.

4.2 Antonyms

Antonym represents opposition of meanings [4]. The words are antonyms if they are opposites in their meanings. Antonym 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 ਧਰਤੀ has antonyms ਆਸਮਾਨ, ਗਗਨ, ਆਕਾਸ਼, ਅਰਸ਼ etc.

4.2.1 Words which have antonyms with more than one context

Like synonyms, in case of antonyms also there exist some words which have antonyms under more than one context. Following are some words examples of this category.

- The word ਪੱਕਾ has antonym ਕੱਚਾ. Same word has antonym ਅੱਲਾ under another context.
- One more example of this category with word ਫਿੱਕਾ(light) which has antonym ਗੂੜਾ (dark) under one context and ਮਸਾਲੇਦਾਰ (bland) under another context.

5. PHASES OF PUNJABI THESAURUS

To provide list of synonyms and antonyms to highlighted word by user, the thesaurus has to passes through various phases. The first step is to collect those words which are synonyms of each other. Because not so much work on thesaurus has been done yet for Indian languages, this is the big challenge to face to collect huge collection of words along with their synonyms and antonyms from various sources. Once words collected, need to choose right structure to save those words. There are various forms to save such data. While time of saving, it is also important to give attention to those words which have same letters but different meaning and words exist with more than one context. Once database made, next phase to choose programming language by which words can retrieve along with their synonyms and antonyms from database and provides output to user. Last phase of thesaurus is replacing selected word from suggestion list of words with input word.

6. CONCLUSION

In this paper, we have discussed the origin and symbols of Punjabi language and role of thesaurus in the NLP. Various relations used in thesaurus like synonym, autonym *etc.* are also discussed with respect to Punjabi. In the last, elaborated phases of Punjabi thesaurus also.

REFERENCES

[1] K. Narayana Murthy, *"On Automatic Construction of a Thesaurus",proceedings of ICSLT-O-COCOSDA 2004 International Conference*, Vol-1, pp 191-194, 17-19 November 2004, New Delhi

[2] <http://www.krysstal.com/langfams_indoeuro.html> accessed on 15may, 2011

[3] Rupinder Kaur, R.K.Sharma, Suman Preet and Parteek Bhatia, 2010, *PUNJABI WORDNET RELATIONS AND CATEGORIZAT-ION OF SYNSETS*, Thapar University, Patiala.

[4] <http://tdil.mit.gov.in/GurmukhiScriptDetailsApr02.pdf> accessed on 20may, 2011

[5] A.Kilgarriff "*Thesauruses for Natural Language Processing"* published in Proceedings of NLP-KE, Beijing, China, pp.513, 2003.

[6] < http://censusindia.gov.in/Cesnus_Data_2001.htm> accessed on 1march, 2011

[7] Meenu Bhagat, 2007, *Spelling Error Pattern Analysis of Punjabi Typed Text, M.E Thesis*, Thapar University, Patiala.

[8] Rupinderdeep Kaur, 2010, *Spell Checker for Gurmukhi Script, M.E Thesis*, Thapar University, Patial a.