xIlkogretim Online - Elementary Education Online,2021; Vol 20 (Issue 5): pp.1844-1849
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.01.196

A Corpus-Based Analysis of The Lexico-Semantic Relationships of Verbs Used in Saraiki language Newspaper

Noreen Zamir, Visiting lecturer, University of Sahiwal, Pakistan, noreenzameer@uosahiwal.edu.pk
Jahanzeb Jahan, Lecturer, Department of English, University of Education, Lahore.
Bilal Asmat Cheema, Lecturer, Division of Arts & social sciences, University of Education, Lahore.
Kiran Jahanzeb, University of Education, Lahore.

Abstract - The present study intends to develop lexico-semantic relationships among Saraiki verbs. For this purpose, the semantic relationships of the English WordNet model established by Miller (1990) at Princeton University is considered as framework of the study. This study aims at promoting the Saraiki language and is a step towards establishing Saraiki WordNet in the upcoming years. A corpus of 1 million words were created from the Saraiki newspaper Jhoke and after text annotation and encoding, a list of 160 verbs were generated. The lexico-semantic relationships of verbs were developed with the help of machine-readable dictionary and the frequency of each relationship was found using Laurence Anthony's Antconc 3.5.7 version. Results exhibited four semantic relations such as synonymy, antonymy entailment, and troponymywith a different frequency and percentage, essential for developing a WordNet. The lexico-semantic relationships of Saraiki can be helpful in developing digital applications and promotingSaraiki language.

Keywords: **Saraiki language, Lexico-semantics, Corpus linguistics, Verbs**

I. Introduction

The background of the Saraiki language is traced back to the Sanskrit language. The traditional linguists claimed that the language spoken by the people in the region of the sub-continent was not the Sanskrit language (Juke, 1900). The documents from history provide clear pieces of evidence that during the first and second phases of progress, people who lived in this part of the world did not speak Sanskrit. They started to speak Sanskrit during the third and fourth periods of resettlement of Aryan civilization in this region. In history, that period is known as the age of Asoka and at that time two languages, including Pali and Sanskrit were spoken and most of the literature was also composed in these languages. Several linguists have defined the region of the Saraiki language differently and the research on the Saraiki language started almost a century ago. The borders of the Saraiki language; in the west, the Saraiki region has a boundary with the Balochi language, boundaries with Rajput Ana's Hindi dialects in the east, and in the south has a boundary with the Sindhi language, nevertheless, in the north, it is difficult for the researchers to determine to knowwhere the origin of Saraiki as language originated (Brien, 1988).

At the current position, the Saraiki region is the center of Pakistan, with 280N to 330N longitude, as it has its area across both sides of the Indus. This region has areas including Chenab, Sutlej, and till to the region of Northern Punjab. This language has its limits to the Iranian, Balochi, and Pashto languages. In the western part, it has boundaries with Indo-Aryan languages and also enters into the south before Rajasthan Marwari dialects. The border of this language with Sindhi has less clarification and similarly in the east, between Saraiki and Punjabi, there is no certain boundary (Shackle, 1976). There are 26 million speakers in Pakistan whose first language is Saraiki (Census, 2017).

Unfortunately, in the recent past, no special attention is given to the Saraiki language, and if we look at digital aspects of this language, although it has few digital resources in the form of its literature, still technologically it is a very undeveloped language. With the advancement of technology, noticeable progress is made in the area of natural language processing (NLP) Machine learning has modified everything as machines settle human language by recognizing them. The concept of machine learning/ artificial intelligence was invented by linguists while performing psychological experiments. WordNet is one such form of machine learning and WordNet of various languages have been designed and these are stored in a central database system (Vossen, 1998). These are interconnected to Princeton WordNet, although each of these WordNet reveals the individuality of its mother tongue and its structural components. Svensen (2009) discusses the important relations among the lexical items while constructing the WordNet.

From a lexicographical perspective, lexical connections played a key role in developing the WordNet and a special electronic dictionary of any language. This fact is also stated by structuralists about designing the dictionary of any language or particular vocabulary scheme (McCarthy, 2003). Prior studies (Gross, Fischer & Miller, 1989; Miller, 1998) have mentioned that WordNet explicitly helps to distinguish among various syntactic categories of and arranges them semantically as nouns are configured as topical hierarchical structures in the development of linguistic memory, while adjectives are ordered mostly on the principle of antonym associations and placed as N-dimensional hyperspaces. According to (Vider&Orav, 2002), Synset of verbs are connections, and these lie to the base of the trees and are arranged based on relationships between the associations.

WordNet is considered a small piece of a complicated classification in which terms are based primarily on the meaning of words instead of just types of words. Besides, it does not represent semantic associations Miller (1998). These connections are incorporated into the nouns, adjectives, verbs, and adverbs (Miller et al, 1993). Today, technology has progressed and computer systems now interpret and also transcribe one language into the other. The desired features are retrieved for the texts of different languages and voice can also be identified in the digital form.

However, some regional languages and dialects are neglected by the linguists and scholars of other domains. Saraiki language is one such language and now needs some special attention towards converting its literature in a digital format. If no considerable attention is given to the Saraiki language, shortly this language would be part of history. The present research is essential not only to digitize and promote the Saraiki language but also is an attempt towards constructing a Saraiki WordNet. It will give an insight towards saving the Saraiki literary works and will also contribute in developing the different tools and resources that will uphold the status of Saraiki language.

Objective of the study

The major intention of the study is to find out the lexico-semantic relationships of verbs found in Saraiki daily newspaper Jhoke.

Significance of the study

The study will prove useful in developing different web applications for the Saraiki language such as WordNet databases and the will further aid in the development of parts of speech taggers, morphological parsers and other Natural Language Processing activities that can also be carried out from the specialized corpus design. It will also help to maintain the status of Saraiki as a language and will enhancing the confidence of native speakers of Saraiki language.

II. Literature review

The Saraiki language is spoken by almost 10% of the total population in Pakistan (Samina et al., 2020), while around 78,000 people in India use the Saraiki language (Shackle, 1976). The Saraiki language is mainly spoken in diverse regions of Pakistan and even in North India. The Saraiki language is also spoken by people who emigrated from North India to another region of India. This language is similar to Punjabi and Sindhi and has many dialects like Multani (being the most commonly found dialect), Riasati, Jafri, Hindki, and Thali. In the year 2007, based on the percentage-fraction of the world population of native speakers, the Saraiki language was ranked 60th out of 100 languages spoken world-wide (Duong, 2017).

In the last two decades, in particular, corpus linguistics has given the scientific study of language a major boost, and a substantial turn-around. As a consequence of corpus linguistics, not only are researchers now able to analyze, with reasonable simplicity, texts that stretch through millions of characters, they have also become conscious of the interesting observations that can be gained through the application of corpus approaches to textual analyses: observations that have been missing in a human-based study. According to Baker (2010, p. 93), “corpus linguistics is an increasingly popular field of linguistics which involves the analysis of (usually) very large collections of electronically stored texts, aided by computer software”. Corpus linguistics is also a technique or strategy used to analyze linguistic anomalies rather than a linguistic sub-field, similar to fields such as syntax, semantics, sociolinguistics and forensic linguistics.

Newspaper articles are a significant subgenre of the newspaper category and have a great deal of importance in news debate. Headlines are given various roles because they are considered as the starting point of the whole corresponding text. The stint arrangement discusses the process of verbal representation including a focus on lexical artifacts of a language and reliance of these artifacts on grammar (Hunston& Francis 1998, 1999). As stated by and Hunston Mason and (2004), Patterns are a set of items in which a word course, category, clauses, or lexical objects are part of each unit. The use of lexical items is considered as the major language feature of newspaper headlines. As asserted by Morley (1998), Headline language can be uncommon, dramatic, and limited. The feature of the vocabulary used in the media is a separate registry.

Verbs may be defined most comprehensively from the vocabulary of all groups. Verb trends are defined by the potential activation of a verb. This approach to grammatical verbs varies from functional analysis to classify the topic, the object, and the balance clause part (e.g., Quirk et al. 1996; Karrlson et al. 1995) or applicant part or event (Fillmore, 1969; Halliday, 1994). Hunston and Mason (2004) had pronounced several verb configurations in three clusters. The first cluster encompasses of the forms which contain a clause component. These patterns are shown in the following example:

- Verb followed by ‘that clause’
- Verb followed by ‘noun group’ followed by ‘wh-clause’

The second group entails the designs which comprise of a single or more than one-word class or group elements. Such patterns are shown in the following examples;

- Verb and noun cluster
- Verb and noun followed by an adjectival group
- Verb followed by an adverb

The third category of verb forms comprises single or more different lexical objects. As depicted in the following instances;

- verb + conjunctions (linking words)
- Verb + possessive + way + prepositional phrase or adverb.

A study was conducted by Mason and Hunston (2004) on the identification of verb forms. To this end, these investigators have used 100 instances taken from the bank of English Corpus' term 'decide.' This research picked the study of Sinclair (1995) to describe the characteristics of the selected verb patterns. And often took a sequential solution to the forms of the chosen verb instead of a hierarchical one. One of the studies carried out by Moe (2014) investigating the articles of media and newspapers published on daily basis to check the language used in them. Thirty-one newspapers were gathered for the study in an effort of making it comprehensive. This thesis analyzed the vocabulary used in newspapers at the textual, graph logical, syntax, conceptual and lexical levels. Kan and Klavans (2015) asserted that an awareness of the application and frequency of verbs will provide valuable insight into the substance and form of an essay. Such researchers often suggest that verbs are a vital part of a sentence and can provide an insight understanding the examining a theoretical and conceptual plot for different items and measures of a text. Often, verbs may assist in classifying publications into various categories. Biber (1989) in his study suggested verbs in 3 categories: social, public, and suasive. (Znamenskaya, 2005) identifies lexical change and syntactic variability contained in the articles of newspapers. It is claimed in her study, the lack of supports, verbs and supporting verbs, the nominalization, and the usage of complicated noun sentences, gags, and the use of tiny rapports are vocabulary characteristics that spot the headlines of a newspaper. A study by (Mouzuaityte, 2015) reviewed Britain media articles to test media form, evaluate newspaper headline vocabulary characteristics, and show the number of stylistic features included in newspaper headlines. As the quantitative methodology, the writer has used detailed mathematical and analytical conceptual analysis. Some of the distinctive characteristics of newspaper headlines are the absence of relative pronouns, auxiliaries, determiners, articles, verbs, and names, according to the results of this report.

WordNet is an on-line lexical database framework whose architecture is influenced by modern human lexical memory psycholinguistic hypotheses. English nouns, verbs, and adjectives are grouped into synonym groups, each reflecting a lexical meaning that underlies them. Different relations link the synonym sets. Verbs probably can be considered the utmost essential category of a language in terms of syntax and lexicons. Both English phrases will have a minimum of one adjective, although as shown in structural phrases, which use "replica" as the subjects of their sentence such as Snowing Show, such phrases do not need to include a linking noun. The sentence paradigm was proposed by several linguists suggesting that verbs hold the central role and stand as the fundamental part of sentences. (Fillmore, 1968; Chafe, 1970). The verb sets out a framework that works for semantic & relational aspects of the sentence. The framework which argues about predicate determines the potential syntactic form of the clauses and sentences where they can appear. The nouns often claim in terms of linking to their positions or instances, such as Tool, specifies the various definitions of the occurrences or situations symbolized by the statement, and the collection constraints define the nouns as per their semantic categories that can form the picture. So, the material comes under the umbrella of semantics and syntactic expression is usually considered as an entity that shows the lexical fragment of the verb, which is, a chunk of the knowledge regarding the verb that is contained in the internal lexicon of the speaker. Due to the difficulty of this material, verbs are perhaps the most difficult to learn lexical type.

The Verb Net (Kipper-Schuler, 2005) classes, reflect clear natural clusters of verbs with common semantic and syntactic properties, arranged in a shallow hierarchy of categories (herein named super classes), groups and subclasses (if any). Super classes combine groups linked to a single form of eventuality, e.g., 'Adding Verbs,' 'Removal Verbs,' which have functional generalizations that are semantically dependent. In addition to the conceptual knowledge obtained from the Frame Net constructs, they are used to facilitate the mapping of Frame Net and WordNet, as well as to further overcome confusion and discrepancies throughout the grouping. Although the hypernymy / hyponymy relationships that include the essential hierarchical structure of verbs and nouns in word Net are clear, membership of the synset to the hypernym tree just provides a very general understanding of the semantic class of which the synset members belong.

WordNet is called a tiny portion of a complex grouping that focuses definitions mainly on the definition of words rather than only forms of words. Therefore, it does not reflect Miller's (1998) semantic relations. These relations are inserted into the terms, adjectives, verbs, and adverbs (Miller et al, 1993). Nowadays, technology has evolved and neural networks are already translating and transcribing one language into another as well. The appropriate features are extracted for text in different dialects, languages and the speech may also be described in a digital type.

Although, there are also ethnic languages and dialects ignored by linguists and academically in other cultures. The Saraiki language is still overlooked by study and now needs to be given particular consideration to translating its literature into a digital format. Unless no significant consideration is given to the Saraiki language, the language will in the immediate future be part of history. As a linguist, it is my major choice to raise some measures that can help to save and promote the different languages. The new work is very important not just for digitizing and spreading the Saraiki language but also for developing the WordNet. It will provide insight into preserving the Saraiki literary works and also lead to the creation of the numerous tools and resources that will maintain the status of the Saraiki language.

III. METHODOLOGY

This study utilizes a quantitative corpus-based approach (Bennet, 2010). The main objective of the study was to develop the lexico-semantic relationship of verbs in the Saraiki language therefore, the researchers decided to create a real-world corpus.

Corpus development

The collection of metadata is the first step in the creation of a corpus therefore, the sample for the corpus development was Saraiki language newspaper 'Jhoke' being published from Multan and was selected purposively in the BO1 file type from July 2020 to December 2020. A specialized corpus of 1 million words was developed from the Saraiki newspaper.

Text processing

The Saraiki corpus was encoded with UTF-8 encoding. This was done to save the corpus in a notepad file so that the corpus can be annotated with a POS tagger to generate the list of verbs.

Corpus Analysis

After text annotation list of 160 verbs were generated and the semantic relations for Saraiki verbs were found manually by looking at the semantic relation of each Saraiki verb using a machine-readable dictionary of Saraiki language. This online dictionary takes input in English and gives output in the Saraiki language. The possible relationships for each verb were also analyzed using the Saraiki dictionary published by Jhoke publishers, Multan (Marziarz, 2011). The semantic relations of Saraiki verbs were verified from a native Saraiki speaker and the frequency of each semantic relation was found using Antconc 3.5.7 version.

IV. Results and Discussion

The Princeton WordNet exhibits four lexico-semantic relations for verbs which include Synonymy, Antonymy, Entailment, and Toponymy. The frequency and percentage of each lexico-sematic relationships for 160 Saraiki verbs are presented below.

4.1 Relationship of synonymy

Table 1 shows the frequency and percentage of the relationship of synonymy of Saraiki verbs

Semantic Relationship	Frequency	Percentage
Synonymy	82	49.39%

Table 1 illustrates the frequency and percentage of the relationship of synonymy found among the verbs of Jhoke newspaper. The frequency of verbs from the corpus that showed synonyms were 49.38% with a frequency of 82 verbs. The remaining 84 verbs did not show any synonym.

Synonyms as defined by Miller (1990) inculcate verbs that are changed or replaced with respective verbs but after changing they do not lose their context and meaning. Some of the verbs which could not show the relation of synonymy include,پکڑن, ڈرنڑ, ڈسنڑیجے ھ بہ, etc

4.2 Relationship of Antonymy

The relationship of antonymy plays an imperative role in lexical databases and WordNet exhibits four different kinds of autonomous relation among verbs which include Reversive opposites, gradable opposites, complementary opposites, and near opposites. The opposition relations for the WordNet are psychologically salient not only for adjectives but also for verbs. The lexico-semantic relationship of verbs is more complex than that of adjectives and nouns (Miller, 1990).

The frequency and percentage of these relationships are presented in the table below.

Table 2 Relationship of antonymy

Lexico-semantic relationship	Frequency	Percentage
Gradable opposites	56	33.73%
Complementary opposites	89	53.61%
Reversive opposites	31	18.67%
Near opposites	22	13.25%

Table 2 shows the types of relationships of antonymy found among the verbs of Saraiki newspaper Jhoke.The complementary opposites were found to be in higher percentage among all the types of antonyms and their frequency came out to be 89. The gradable opposites which are measured or found on the intensity of grades were 33.73% of the total list of nouns. Whereas, the reversive and near opposites were less frequent among the verbs and were 32% of the entire list.

4.3 Relationship of Entailment

The relationship of entailment is a significant element of verbs in WordNet. Many verb pairs in an opposition relation also share an entailed verb. Owing to the bidirectional relation the verbs share two terms such as entailing and entailed (Chklovski&Pantel, 2004).

The table below shows the frequency and percentage of entailing and entailed verbs in the corpus.

Table 3 Frequency and percentage of the relationship of entailment

Lexico-semantic relationship	Frequency	Percentage
Entailing	89	53.61%
Entailed	56	33.73%

4.4 Relationship of Troponymy

The principal relation linking verbs in a semantic network is the manner relation (or “troponymy”).Troponymy is not a semantically homogeneous relation; rather, it is polysemous and encompasses distinct sub-relations (Khokhlova, 2014). Troponymy is not a semantically homogeneous relation; rather, it is polysemous and encompasses distinct sub-relations. The verbs Jhoke newspaper corpus showed the dual processes and the table below shows the frequency and percentage of verbs showing troponymy.

Table 4: Frequency and percentage of the relationship of Troponymy

Lexico-semantic relationship	Frequency	Percentage
Troponymy	110	66.26%

Table 4 shows that out of 166 verbs 110 Saraiki verbs showed the relationship of troponymy and their percentage came out to be 66.26%. The verbs of newspaper Jhoke such as ہرݨ,چلاوݨ, دھوݨ,etc showed the dual-process i-e relationship of toponymy.

V. Limitations

The objectives and design of this research were subject to a few limitations. The major limitation which the study faced regarding the instruments and techniques was to save the Saraiki script in a notepad file as, the corpora are saved in a text file so that they could be processed with Natural language processing software. To address this limitation the corpus was encoded with UTF-8 encoding in order to save the corpus in a notepad file.

VI. Conclusion and future research

The Results revealed four semantic relations among the Saraiki verbs, essential for the development of a WordNet. These lexico-semantic relationships include synonymy, antonymy entailment, and troponymy and all the four semantic relationships showed different frequency and percentages. The relationship of antonymy was found to be most frequent among all the semantic nets.Lexico-semantic relationships of a language are a key design for the construction of a digital thesaurus. Ever since the researchers have started working on Natural language processing the linking of lexical items have become essential characteristic of electronic dictionaries and other computational tools. Without a specialized corpus it becomes nearly impossible to develop digital applications of a language.

Moreover, Jhoke newspaper corpus can also be utilized for developing Natural language processing tools such as POS taggers, psyche taggers, named entity recognition and morphological parsers.

References

1. Bennet, R.G. (2010). Using corpora in the language learning classroom: Corpus linguistics for teachers. Michigan ELT, 34-48.
2. C. Shakle (1976). The Saraiki Language of Central Pakistan, p.1.
3. Gross, D., Fischer, U., & Miller, G. A. (1989). The organization of adjectival meanings. Journal of memory and language, 28(1), 92.
4. Jukes Dictionary of Jatki or western Punjabi. 1900. Preface.
5. Khokhlova, L. V. (2014). 2. Majority Language Death. University Of Hawai'i Press.
6. McCarthy, D., Keller, B., & Carroll, J. (2003, July). Detecting a continuum of compositionality in phrasal verbs. In Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18 (pp. 73-80). Association for Computational Linguistics.
7. Miller, G. (1998). WordNet: An electronic lexical database. MIT press.
8. Miller, G. A. (1990). Nouns in WordNet: a lexical inheritance system. International journal of Lexicography, 3(4), 245-264.
9. Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T. (1993, March). A semantic concordance. In Proceedings of the workshop on Human Language Technology (pp. 303-308). Association for Computational Linguistics.
10. O. Brien. (1988). A Glossary of the Multan Language. Punjab Government Press, p.1.
11. Svensen, B. (2009). A handbook of lexicography: The theory and practice of dictionary making. Cambridge: Cambridge University Press.
12. Samina. K., Uzma. N., Muhammad. I., & Younas. M., (2020). THE STUDY OF ORTHOGRAPHICAL DIFFERENCE BETWEEN PUNJABI LANGUAGE AND SIRAIKI DIALECT IN PUNJAB PROVINCE. HamdardIslamicus, 43, 175-193.
13. Vider, K., &Orav, H. (2002). Estonian wordnet and lexicography. na.
14. Vossen, P. (1998). A multilingual database with lexical semantic networks. Dordrecht: Kluwer Academic Publishers. doi, 10, 978-94.