Research Journal of Language and Literary Studies
Spring 2024, Volume 4, No.1, pp 17-50

ALIGNING ENGLISH AND PUNJABI WORDNETS: A CORPUS-BASED STUDY

Muhammad Farukh Arslan[1], Prof. Dr. M. Asim Mahmood[2], Hira Haroon[3]

ABSTRACT

The Punjabi language is primarily spoken in the Punjabi region of the subcontinent. In Pakistan, the variant spoken is known as Shahmukhi Punjabi, and it is written in the Perso-Arabic script, while in India, it is written in the Gurmukhi script. Punjabi serves as the mother tongue for 38.77% of Pakistan's population. Despite its historical significance, literary wealth, and a large speaker base, Punjabi has not attained the status of a prestigious, digitally advanced language. To address this, our study aimed to develop a Punjabi WordNet and align it with the English WordNet. The analysis process involved collecting a corpus of nine million (9M) words, cleaning and tagging it to extract grammatical categories, and morphologically analyzing vocabulary items to understand word formation patterns, placing them under specific grammatical tags. The morphological analysis was conducted using Distributed Morphology (DM), revealing the patterns of combining derivational and inflectional morphemes with the root. During this research, a variety of software and digital resources were developed. These include a spell checker, a morphological analyzer, a parts of speech (POS) tagger, a semantic tagger, the Punjabi WordNet, and an alignment tool.

Keywords: Alignment, English and Punjabi Wordnet, Corpus based study

[1] Lecturer NUML Fsd campus, PhD scholar GCUF.
[2] Dean Faculty of Arts and Social Sciences, GCUF
[3] Lecturer GCUF.

1. Introduction

This study is about the alignment of the WordNet of the Punjabi language with the Princeton English WordNet. The WordNet is a lexical resource in which words are linked to each other on the basis of associations of meanings and forms (Fellbaum, 1998). WordNet is a lexical database that provides nouns, verbs, adjectives, and adverbs that are grouped into sets of cognitive synonyms (i.e., synsets), each expressing a distinct concept. The synsets are interlinked by means of conceptual-semantic and lexical relations.

WordNet is a large lexical resource which can assist us in word sense disambiguation, dictionary development, vocabulary learning, machine translation, artificial intelligence and evaluation (Capel, 2012). Nation (1990) declared WordNet as a key source for the task of language teaching and learning.

Punjabi language is a member of Indo-Aryan branch of the Indo-European language family. It is a historically rich language and can be traced back to the inhabitants of Indus Valley at least as far 2500 BC (Rahman, 2002). Punjabi language is considered as the 10th widely spoken language in the world (Ghai & Singh, 2013) and Punjabi language is the native language of 125 million people across the globe. Therefore, the importance of Punjabi language is an exceptional phenomenon.

Arslan, Mehmood and Hayat (2019) developed a vocabulary profile of Shahmukhi Punjabi language. The frequency list and list of different vocabulary items according to their grammatical categories were studied in the developed corpus. It has been observed that the words of Punjabi language have many different cases and forms as contrary to English language and similar to Urdu language. Nouns, verbs and adjectives vary according to number and gender. Abbreviations and loan words from English language were also found in the corpus. It was an impressive attempt to highlight Shahmukhi Punjabi and raise its status.

The Punjabi language is rich in linguistic patterns as well as literature, but it has been neglected by the research community and lack of research is conducted on Punjabi language. Moreover, it is surprising that no significant work has been done on the lexical resources and digitization of this language. The present research is a step to create digital lexical resources of this language and develop an alignment with English language WordNet on the basis of a special purpose corpus, and it aims at the alignment of Punjabi and English WordNet in terms of synsets.

1.1 Implication of the Study

The development of Punjabi WordNet and its alignment with English WordNet will play a major role in making it digitally advanced, and it will fill the gap regarding the linguistic resources of Punjabi. All the utilities and linguistic stock will also assist in saving the language from extinction. Arslan and Mahmood (2021) reported about the sound shift and sound loss in Punjabi. Speakers of Punjabi are shifting to Urdu and English at a very high rate, as a result of which sound shift and sound loss are occurring very frequently. So, the development of Punjabi WordNet and its alignment with English WordNet is the need of the hour. It will also improve the status of the Punjabi language and save it from extinction by developing its linguistic resources.

1.2 Significance of the Study

This study is crucial and has direct influence in the areas of NLP, CL, sociolinguistics, and language teaching and learning. The development of WordNet will affect the tasks of NLP and CL, and it will help in the development of other utilities in the area of NLP such as morphological analyzer, POS and semantic taggers. WordNet will assist in the tasks of artificial intelligence, language digitization, software development, and online language teaching. 5 This is the era of digitization and online teaching. It is impossible to proceed with online teaching without developing the lexical database of Punjabi. WordNet will provide the lexical database

necessary for online teaching, and it will help in the development of online dictionaries of Punjabi language and take Punjabi language to a higher and advanced platform.

1.3 Research Questions

Q 1. What are the morpholigcal patterns of Punjabi lexical items?

Q 2. What are the different lexico-semantic relations of nouns, verbs, adjectives, and adverbs in Punjabi WordNet?

Q 3. What is the process of developing synsets in Punjabi WordNet?

Q 4. What kinds of similarities and differences are found among the synsets of Punjabi and English wordnets?

Q 5. How is the alignment made between Punjabi and English wordnets?

2. Literature Review

In this section of the study, initially the area of natural language processing (NLP) is discussed secondly, wordnets are discussed mainly focusing on the methodology for the development of wordnet. Further, different the wordnets developed in the past are mentioned. Just to highlight the importance and implications of the wordnets. Lastly, the studies related to the mapping the wordnets have been mentioned to strengthen the claim and need of the study and develop Punjabi wordnet and align it with the English wordnet.

The discipline of NLP emerged in the 1950s. It combines the knowledge of computers and languages for better efficiency, working, and understanding. NLP provides the algorithms and systems for a better understanding of language (Rospocher et al., 2016). NLP has two basic characteristics: natural language understanding and natural language generation (Khurana, Koli, Khatter, & Singh, 2017).

Research works in the domain of NLP are a continuing trend, and many works are found in the domain of NLP with relation to POS tagging, sentiment analysis, emotion detection, and named entity recognition (NER)

(Jeonghee et al., 2003). Applications of NLP include machine translation, which is translating phrases from one language to another, such as with Google Translate. It employs various techniques; examples of such techniques are word error rate, position-independent word error rate (Tillmann et al., 1997), generation string accuracy (Bangalore et al., 2000), multi-reference word error rate 9 (Nießen et al., 2000), BLEU score (Papineni et al., 2002), and NIST score (Doddington, 2002).

Since the 1990s, Wordnets have continuously been built for more than 150 languages around the world; many of those Wordnets are typologically and genetically unrelated to the original English WordNet (Bosch & Griesel, 2017). The importance and implication of the wordnet have been identified very efficiently, which led to the development of Wordnets in so many languages.

Putra, Arfan and Manurung (2008) termed two basic approaches for the development of WordNet: the expansion approach and merge approach. These two approaches can be followed in the development of WordNet either the merge approach or the expansion approach (Vossen, 1998). First, merge approach basically follows the rule in which all the possible senses of the word are collected and recorded at first stage (Puspak, 2010). At the second stage lexicographer develops a synset for each of the senses. Synset development is obeyed following the three principles of minimality, coverage and replicability and then at the next stage synsets developed in that language with Princeton WordNet (Kulkarni, Dangarikar, Kulkarni, Nanda, & Bhattacharyya, 2010). In merge approach no other language influences and distracts in the process of synset development. Lexicographer plays a vital role by incorporating the cultural and regional concepts of the source language. Therefore, the merge approach is time consuming (Akhter, Mahmood, & Nadeem, 2019).

Princeton University made the first ever attempt regarding the development of English WordNet in 1990 by Miller (1995). It consisted of

155,000 synsets, as cited in Fellbaum (1998). English WordNet is an open-source utility that is easily accessible. The development of English WordNet gave a new direction in the fields of NLP and CL.

Bosch and Griesel (2017) conducted the study regarding the development of WordNet for under-developed languages. African languages were the major concern of the study. The expansion model was used in the study of data analysis. In this regard, Ordan and Wintner (2007) and Vossen (1998) debated two methods for the development of WordNet. Moreover, Felbaum (1998) elaborated on the "merge approach" as the alignment of the newly developed WordNet with the Princeton WordNet. On the other hand, the expansion approach follows the approach of translating the indigenous words directly to English WordNet tools by occupying the available dictionaries (Zafar, 2012). Lastly, Bosch and Griesel (2017) developed the WordNet of African languages and aligned it with the English WordNet.

Putra, Arfan and Manurung (2008) reported the development of web-based annotation tool. It was termed as LAMP stack which was based on the MySQL DBMS backend. Web scripting language which was employed in the study was PHP. There were few sections reported in the study; user management assisted in automatic assignment of the synsets which were mapped. It provided the facility to look for the word senses. The last constituent was reporting of statistics, such as number of users, number of synsets mapped and the facility to extract the mappings for further analysis.

Lee and Lee (2000) reported the construction of the Korean WordNet form Pre-existing lexical resources. Word Sense disambiguation techniques were implemented to link Korean Words with English WordNet synsets. Machine learning methods were used to combine the word sense disambiguation techniques.

Ramanand, Ukey, Singh and Bhattacharyya (2007) claimed the observations on the structural properties of the WordNets of three languages: Hindi, English and Marathi. Study reported work on linking Hindi, English

and Marathi synsets. Study offered a formula for calculating the similarities of the nodes in two hierarchies.

Farreres (2005) proposed a methodology for mapping the Spanish thesaurus to English WordNet consisting of two steps. The methodology is organized as order of two processes. The Purpose of the basic process which is based on this work is mapping of Spanish words to WordNet synsets (Atserias, Climent, Farreres, Rigau, & Rodriguez, 2000). The second process follows the pattern of hierarchies with relation to the associations produced in the first phase.

Princeton University pioneered the development of English WordNet in 1990, as initiated by Miller (1995). It encompassed 155,000 synsets, as documented by Fellbaum (1998). English WordNet stands as an open-source tool, easily accessible and transformative in the fields of NLP and Computational Linguistics. Following this landmark, the development of WordNets continued worldwide:

♣ Spanish WordNet, established in 1999, boasts 62,000 synsets and is readily available for research purposes (Farreres, Rigau & Rodriguez, 1998).

♣ Russian WordNet, introduced in 2004, incorporates 117,000 synsets and operates as an open-source utility (Balkova, Suhonogov, & Yablonsky, 2008).

♣ Thai WordNet, created in 2007, includes 77,593 Synsets and is accessible as an open-source WordNet (Thoongsup et al., 2009).

♣ Japanese WordNet, developed in 2008, features 57,000 synsets and is accessible as an open-source utility (Isahara et al., 2008).

♣ Arabic WordNet, also established in 2008, consists of 11,269 synsets (Black et al., 2006). Numerous other WordNets are available online.

While this provides an overview of WordNets developed in the past, it underscores the research gap that exists, particularly the absence of a Punjabi language WordNet, despite the presence of WordNet for many

European and some Asian languages, including Hindi and Tamil. Developing a WordNet for Punjabi language remains an important endeavor.

Mandala et al. (1998) proposed using WordNet as a source for the automatic construction of thesauri. Moldovan and Mihalcea (2000) highlighted WordNet's role in developing natural language interfaces to enhance the accuracy of internet search engines by expanding queries. WordNet provides lexical relations and aids in discerning word meanings within context (Towell & Voorhees, 1998). It proved to be a valuable source and a key methodological element for word meaning disambiguation and information extraction systems (Chai & Biermann, 1997). GINGER II, a product of English WordNet, played a crucial role, incorporating semantic categories developed within WordNet.

WordNet holds significant implications for language teaching, learning, and translation applications. It was even discussed in an international conference on WordNet in the context of foreign language teaching. Hu, Graesser, and the Tutoring Research Group (1998) suggested using WordNet as a rich source of vocabulary to assess learners' command of the target language's vocabulary. WordNet aids in translation tasks through semantic distance algorithms and assists in placing words more appropriately within context (Shei & Pain, 2000). WordNet also supports inter-linguistic arrangements and underscores the need for expert translation retrieval systems. The CINDOR system, 27 evaluated for cross-linguistic translation systems, proved to be a useful application (Diekema, Oroumchian, Sheridan & Liddy, 1998).

3. Methodology

This section briefly explains the steps followed in to conduct the study. Initially, the corpus collection has been mentioned. Secondly, the extraction of wordlist has been mentioned. Thirdly, the morphological analysis of Punjabi vocabulary items has been mentioned. Fourthly, manual POS tagging and tagger development has been mentioned. Fifthly, synset

development has been mentioned and explained. Lastly, the development of utility and alignment of Punjabi and English wordnets have been mentioned.

3.1 Corpus Collection

The first and foremost step regarding this study was the development of corpus. Nine million (9M) words corpus of Shahmukhi Punjabi was collected. Books, newspapers, novels, short stories and online websites containing the linguistic text of Shahmukhi Punjabi were considered as the basic source of the corpus.

3.2 Wordlist Extraction

The next stage was the extraction of wordlist of the corpus. Wordlist proves as the key element for any research. It justifies the overall division of the vocabulary items covered in the corpus from overall language. The whole corpus was loaded in the Antconc and the wordlist was extracted. Wordlist tells about the overall frequency of the words occurred in the corpus.

3.3 Morphological Analysis

Morphological analysis of Punjabi vocabulary items was conducted. The analysis was performed by using the the theory of DM. This analysis helped in understanding the word formation process of Punjabi language. In result of which, it was easy to place vocabulary items in the list of specific grammatical category.

3.3.1 POS Tagging

Vocabulary items of the corpus were tagged through POS tagger to get specific lists of each grammatical category. POS tagger was developed for tagging in this study. At first step vocabulary items were analyzed manually. While analyzing the morphology lemma words were separated, the inflections were separated and the vocabulary items were separated according to their specific grammatical categories.

3.3.2 Synset Development

In the next stage, synsets of Punjabi words were developed. Synsets consist of a word, its gloss (explanation), synonyms, and examples in the

form of sentences. Vocabulary items were categorized based on their specific grammatical category, such as nouns, verbs, adjectives, and adverbs. Lists of these grammatical categories were separated from each other.

3.3.3 GUI Utility Development

The utility was created to manage synsets and semantic relations through a graphical user interface, constructed using the tkinter library in Python. Essential libraries, including tkinter for GUI components, pandas for data manipulation, nltk for NLP, and other specific libraries, were imported.

The utility defined a class named ExcelViewerApp, which inherited from tk.Tk and functioned as the core of the GUI operations. The __init__ method was employed to initialize the main window, configuring the window title, size, and creating buttons, entry fields, and frames to structure the layout.

In the subsequent phase, the synsets of Punjabi WordNet were aligned with those of Princeton English WordNet. To achieve this, the translation of each word was searched for in Princeton English WordNet. The corresponding gloss and sentences for Punjabi words were retrieved from Princeton and inserted into the Excel file. After this step, every gloss and sentence of a Punjabi word had an equivalent gloss and sentence in English translations. If any word or sense in the Punjabi language lacked a corresponding sense in the English language, the corresponding cells in the English column were left empty.

In the following stage, all the sheets formatted in Excel were integrated into the backend library, and a GUI application was created. This application enabled users to search for a Punjabi word, which would then display the English translation of the Punjabi word, along with the Punjabi gloss and sentences, as well as their English counterparts.

4. Results and Discussion

Comprehensive morphological analysis was conducted to understand the system of singular, plural in noun, masculine and feminine markers in

nouns, agreement between nouns and adjectives, verb froms and patterns and lastly adverb formation in shahmukhi Punjabi. Roots of each word and all its inflectional and derivational forms were extracted with the help of morphological analysis.

Pattern 1: Making masculine plurals with the insertion of /e/ ے

Syntactic Function	**Morphological Function**		**Semantic Function**
Root morphemes	**Inflections**	**Merging**	**Logical form**
√Munda منڈا √Kamra کمرا √Khota کھوتا √ghorra گھوڑا √Kerra کیڑا √Danda ڈنڈا <N°, Gen, Num>	/e/ <Masc, Sing>	Last vowel is substituted by /e/ to make plural	Munde منڈے Kamre کمرے Khote کھوتے Ghorre گھوڑے Kerre کیڑے Dande ڈنڈے <Masc, Plu>

The root words such as √Munda are head word here and it joins the inflectional morpheme /e/ to make plural. The syntactic function here is that the root √Munda has the following abstract features <N°, Gen, Num>. In the morphological functional morpheme is added to the root to make plurals.

Pattern 2: Making plurals with the insertion of /an/ or /ian/

Syntactic Function	Morphological Function		Semantic Function
Root morphemes	Inflections	Merging	Logical form
√Kuri کڑی √Roti روٹی √Kitab کتاب √Pinsal پنسل √keri کیڑی	/an/	/an/ or /ian/ are added at the end to make plural	Kurian کڑیاں Rotian روٹیاں Kitaban کتاباں Pinsalan پنسلاں Kerian کیڑیاں

<N°, Gen, Num>	<Fem, Sing>		<Fem, Plu>

Feminine nouns in Punjabi have roots in the form of √Kuri, and they are marked for number to create plurals in their syntactic function. The abstract description of this root is represented as <N°, Gen, Num>. In the realm of morphological functions, the inflectional morphemes /an/ ان is added as suffixes to form plural nouns from feminine singular nouns. There is a distinction between these two morphemes: for feminine nouns that end with the vowel sound /i/ ی (which is a key indicator of a feminine noun), this vowel persists, and the morpheme /an/ is added to form the plural. However, for feminine singular nouns ending with a consonant sound, such as the root word √Kitab, only the morpheme /an/ ان is added as a suffix to create the plural.

Pattern 3: Making plurals with the insertion of /wan/

Syntactic Function	Morphological Function		Semantic Function
Root morphemes	Inflections	Merging	Logical form
√Dua دعا √Dawa دوا √Saza سزا	/wan/	Feminine nouns ending with vowel morpheme /wan/ is added at the end to make plural	Dua**wan** **دعاوان** Dawa**wan** **دواوان** Saza**wan** **سزاوان**
<N°, Gen, Num>	<Fem, Sing>		<Fem, Plu>

For other feminine nouns ending with the vowel sound /a/, like the root word √Dua, a derivational morpheme VIs /wan/ is used to create plurals.

The abstract description of the root word remains as <N°, Gen, Num>. In the morphological function, the VIs /wan/ واں is attached to the root words to form plurals. This morpheme is specifically added to feminine nouns that end with the /a/ sound. The abstract description of the semantic function after attaching VIs at the suffix position is <Fem, Plu>.

4.1 Lexico-Semantic Relations of Nouns

Here, the lexico-semantic relations of nouns are given in the table below.

NOUNS	**ENGLISH**	**SYNONYMY**	**ANTONYMY**	**MERONYMY**	**HYPONYMY**	**HYPERNYMY**	**HOLONYMY**
بھیڑ Bheer	Crowd	رش Rush	سنسان Sunsan	بندہ Bandah	X	معاشرہ Muashra	X
فرنگی Firangi	Englishman	انگریز Angraiz	پینڈو Pendoo	X	X	انسان Insan	X
بندا Banda	Ear-ring	بالی Baali	چوڑی Choorri	سونا Sona	سونے دا بندا Sonay da banda	زیورات Zevrat	X
بُدّھی Budhi	Wisdom	سیانا Syana	جھلا Jhalla	X	X	انسان Insan	X
پاتشاہ Patshah	king	بادشاہ Badshah	فقیر Faqeer	X	X	انسان Insan	حکومت Hakoomat
بگرا Bakra	Goat	گوسفند Gosfand	مجھ Majh	پوشل Poshal	پاڑی بکرا parri bakra	ڈنگر dangar	X
منڈی Man	market	بازار Bazar	گھر Ghar	دکان Dukan	سبزی منڈی Sabzi	X	X

ddi					Mandi		
بنگلا Bangla	bungalow	حویلی Haveli	چونپڑی Chonprri	کمرہ Kamrah	وڈی حویلی Waddi haveli	رہائش Rihaish	X

In the table above, we have provided examples of lexico-semantic relations specifically pertaining to nouns. Nouns predominantly exhibited synonyms and antonyms for all words. Synonymy, regarded as the most important and widely discussed relation, signifies the connection between words that convey the same meaning (Yule, 2020). In the context of this study, all nouns featured synonyms in Punjabi, which were meticulously developed and added to the database. The second prominent relation was antonymy, which denotes the opposition of meanings (Palmer, 1996). Similar to synonyms, all nouns possessed antonyms, encompassing both direct antonyms and words with contrasting meanings.

However, variations were observed regarding other lexico-semantic relations of nouns, such as meronymy, hyponymy, hypernymy, and holonymy. Meronymy represents a semantic relation that signifies the whole-part relationship between lexical items. In cases where X has Y, Y is considered a part of X (Ullmann, 2019). Variations arose concerning the lexico-semantic relation of meronymy, as not all nouns exhibited this relation. Some lexical items lacked any whole-part relationship with other nouns, such as the word "منکا" (manka) meaning "bead," which did not possess a meronymic relation due to its lack of further parts.

4.1.1 Graphical Representation of Nouns

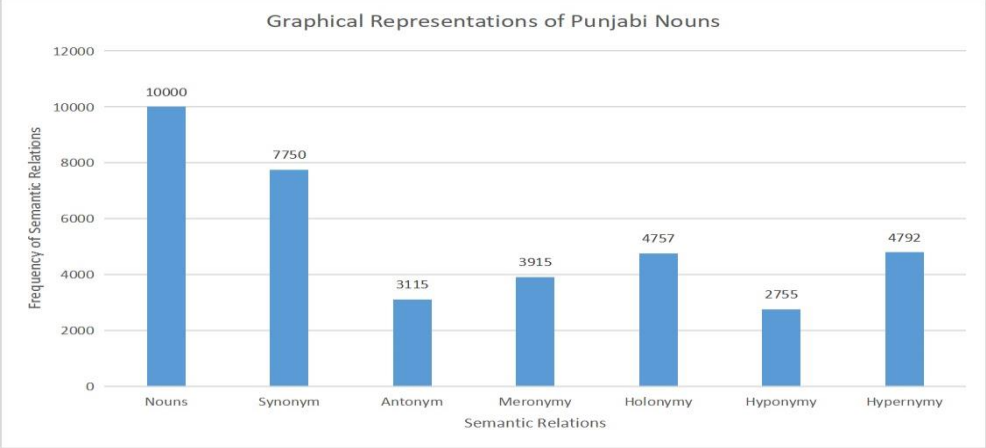


Figure 4.1 Lexico-Semantic Relations of Noun

The analysis showed that there were totaL10,000 nouns taken in this study for the development of synsets and semantic relations. Out of those 10,000 nouns, 7750 nouns possessed synonyms, 3115 antonyms were developed, and 3915 meronyms, 4757 holonyms, 2755 hyponyms and 4792 hypernyms were developed of Punjabi WordNet. The analysis revealed that 77.50 % synonyms of nouns were developed, 31.15 % antonyms, 39.15 % meronyms, 47.57 % holonyms, 27.55 % hyponyms and 47.92 % hypernyms. Such variation in the numbers of lexico-semantic relations appeared due to variation in the semantic categories of the words and the nature of the words. These factors played vital role due to which certain lexico-semantic relations were not developed of certain words.

4.2 Lexico-Semantic Relations of Verbs

Verbs, as action words, have been discussed in detail in the previous section. Verbs were categorized into action verbs and helping verbs, with action verbs further classified as one-word verbs, conjunct verbs, and compound verbs. This study primarily focused on developing lexico-semantic relations for action verbs. Below, we provide examples of these relations:

Conjunct	Transl	SYNONYM	ANTONY	TROPON	ENTAIL

Verb	ation	Y	MY	YM	MENT
جھات پان Jhaat paan	Peek	تک پان/ ویکھن Takk paan/ Wekhan	نظر انداز کرن Nazar andaz karan	ویکھن Wekhan	مشاہدہ کرن Mushahida karan
ونڈ پان Wand paan	Segrega te	جدائی پان Judai paan	ملان/ اکٹھا کرن Milan/ Ikatha karan	علیحدہ کرن Alehda karan	طلاق دین Talaq den
نمو پان Namoo paan	Grow	ودھن Wadhan	گھٹن/ کم ہون Ghattan/ Kam hon	بڑھن/ودھن Barrhan/ wadhan	ترقی کرن Taraqi karan
پردا پان Parda paan	die	مر جان / فوت ہون Mar jan	جینا/ زندہ رہن Jeena/ Zinda rehn	قتل کرن Qatal karan	شہید ہون Shaheed hon
قرار پان Qarar paan	agree	طے پان Teh paan	اختلاف کرن Ikhtilaf karan	اقرار کرن Iqrar karan	یقین رکھن Yaqeen rakhan

As previously mentioned, verbs are the action words in language. In the table below, we present lexico-semantic relations specific to conjunct verbs. Synonyms, which are words with the same meaning, and antonyms, which are words with opposite meanings and senses, are common relations shared by most verbs. However, variations were observed in the case of troponyms and entailment.

Troponyms refer to verbs that have similar meanings to another verb but differ in the intensity or manner of the action performed. A troponymic relation exists when one verb provides more specific information about how an action is performed in comparison to another verb. This lexico-semantic relation was thoroughly examined and developed in this study.

4.2.1 Graph Graphical Representation of Lexico-Semantic Relations of Verbs

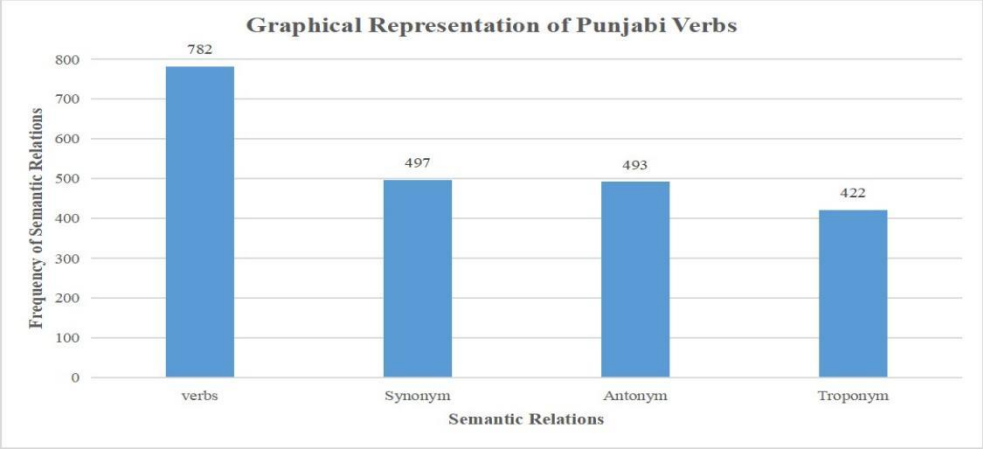


Figure 4.2 Lexico-Semantic relations of Verbs

The graphical representation showed that there was total 782 verbs in corpus. Among these verbs there were 497 synonyms, 493 antonyms and 422 tropoyms. The analysis showed that a significant portion of verbs, 63.55% for synonyms and 63.03% for antonyms and 53.95% of troponyms were successfully developed.

4.3 Lexico-Semantic Relations of Adjectives

Adjectives	Translation	Synonym	Antonym
امیر Ameer	rich	مالدار Maldar	غریب Ghareeb
قابل Qabil	able	لائِق Laieq	نااہل Naehal
درست Darust	accurate	عین مطابق Ain mutabiq	غلط Ghalat
انتظامی Intizami	administrative	ہدایتکاری Hidyatkar	غیر انتظامی Ghair intizami

پسندیدہ Pasandidah	Favorite	من بھاتا Mann bhata	ناپسندیدہ Na pasandeedah

Table 4.5 Lexico-semantic relations of Adjectives

Adjectives play a crucial role in language as they are words that qualify and describe nouns and pronouns. In the table provided above, we present the lexico-semantic relations specific to adjectives. Among adjectives, only two primary lexico-semantic relations were identified: synonyms and antonyms.

4.3.1 Graphical Representation of Lexico-Semantic Relations of Adjectives

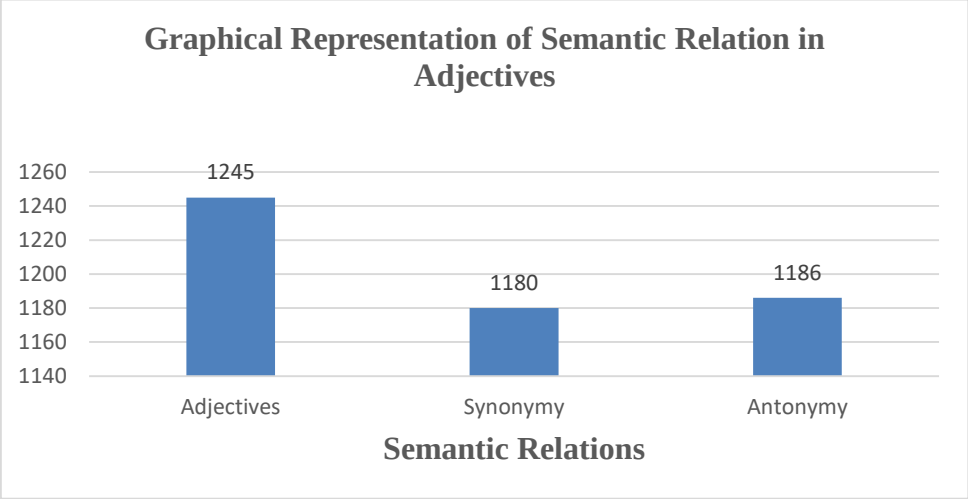


Figure 4.3 Lexico-Semantic relations of Adjectives

The graphical representation of adjective frequencies revealed a total of 1,245 words. Among these adjectives, there were 1,180 synonyms and 1,186 antonyms. These findings indicate that a significant portion of adjectives, approximately 94.78% for synonyms and 95.26% for antonyms, were successfully developed.

4.4 Lexico-Semantic Relations of Adverbs

Adverbs, a vital component of speech, serve the purpose of modifying and qualifying verbs. Within the Punjabi language, the lexico-semantic relations observed among adverbs primarily consist of synonyms and antonyms. Examples of those lexic-semantic relations of the adverbs are given in the table below:

Adverbs	**English Translations**	**Synonymy**	**Antonymy**
دھیرے Dheeray	Slowly	آہستہ آہستہ Ahesta ahesta	جلدی نال Jaldi naal
نہیں پر Nahin par	Else	ہور hor	گھٹ ghatt
نکّو نکّ Nakko nakk	Fully	مکمل طور تے Mukamal tor tay	X
نزدیک Nazdeek	Near	قریب Qareb	دور Dor
نوِت Navit	New	نئی Nai	تےانی Tiani
نا Nah	No	نہیں Nahin	ہاں Haan

Table 4.7 Lexico-semantic relations of Adverbs

The table above presents examples of lexico-semantic relations among adverbs. It's worth noting that adverbs were the only part of speech for which synonyms and antonyms were not universally available for all words. For example, the adverb "دیکھی دیکھا" (dekha dekhi), used to convey imitation, did not have either synonyms or antonyms in the Punjabi language. Similarly, the adverb "نک نکو" (nakko nakk) had a synonym and a lexico-semantic relation but did not possess an antonym.

These instances underscore the distinctive characteristics of Punjabi adverbs and the variability in the presence of synonyms and antonyms among them.

4.4.1 Graphical Representation of Lexico-Semantic Relations of Adverbs

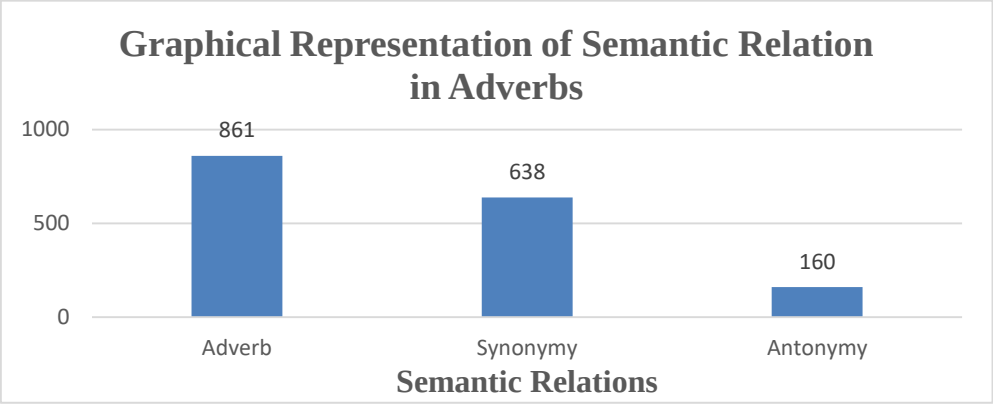


Figure 4.4 Lexico-Semantic Relations of Adverbs

The graphical representation of the lexico-semantic relations of adverbs revealed that there was a total of 861 adverbs. Out of these 861 adverbs, 638 adverbs were associated with synonyms, while only 160 had antonyms as their lexico-semantic relations. This data is summarized in the table below, indicating that 74.10% of adverbs had synonyms, and 18.58% had antonyms as lexico-semantic relations.

4.5 The Process of Development of Synset of Punjabi WordNet

The third research question aimed to describe the process involved in developing synsets for Punjabi WordNet. A synset is a group of words considered synonyms, with similar meanings that can be used interchangeably in specific contexts. Each synset represents a distinct concept or idea, and the words within a synset are related in terms of their semantic meaning. WordNet organizes words into synsets to capture the various senses and meanings a word can have.

A synset in WordNet consists of words that are close synonyms to each other. All the words in a synset share similar concepts, making them close synonyms (Miller, 1995). In this section of the study, the process of developing the synsets is explained. The steps involved in synset development are outlined below:

Development of Corpus: Initially, a corpus of Punjabi language was collected. This corpus included books, newspapers, novels, and other prose written in Punjabi.

Tagging the Corpus: A part-of-speech (POS) tagger, specifically developed for this research, was employed to tag the corpus. After tagging the corpus, lists of the four grammatical categories were extracted based on the tags assigned by the corpus.

4.6 Development of Synset

During the development of synsets, each word was researched in dictionaries to gather all possible senses and contextual meanings. The input and expertise of a native speaker were also sought during the synset development process.

Several observations were made during this alignment process:

1. Variation in Number of Senses: Princeton English WordNet contained a higher number of senses, although many of these senses had very similar meanings and concepts. This discrepancy resulted in differences between Punjabi and English WordNets.

2. Difference in Senses: The gloss found in Princeton English WordNet did not always perfectly match the gloss of Punjabi WordNet. While similarities in sense and context existed, the sense of the gloss was not always identical, and alignment required interpretation of the word's context.

3. Lack of Sentences: It was noted that Princeton English WordNet did not include sentences for all words. During alignment, Punjabi words were translated into English, and the translations were searched in Princeton

WordNet. Alignment was based on the similarity of meaning and concept, but sentences from English WordNet were often missing.

4. Phrase Translation: Punjabi words were translated into English, and these translations were searched in Princeton WordNet. However, Princeton WordNet did not consistently provide results for phrases; it mainly offered results for single words. Even words containing prepositions, articles, and other constituents were often not found in English WordNet.

Examples of phrase translations are provided in the table below.

Table 4.8 Phrase translation of Punjabi words

Word	**English Translation**	**Punjabi Gloss**
رہن سہن Rehan sehan	way of living, mode of life, lifestyle	روزانہ زندگی گزارن دا طور، ڈھنگ Rozana zindagi guzaran da tor, dhang
ونگا Wanga	Rope fastening together the hind legs of a cow while milking	دودھ پلان دے دوران گائے دی پچھلی لتاں نوں جوڑن والی رسی Dudh pilan day doran gaye di pichi latan nun jorran wali rassi
ونگ Wang	Bangle of glass or plastic	شیشے یا پلاسٹک دی چوڑی Shesay ya plastic di chorri
وچار Wichar	the exercise of judgment or reason	فکر، سوچ، غور، تدبر Fikar, soch, ghor, tadabur
واڈھا Wadha	A person employed for reaping crops	فصلاں دی کٹائی لیی کم کرن والا شخص Faslan di katai kam karan wala shakhs
ورد Wird	repeated prayers	کسے گل یا کسے نام نوں بار بار دہران دا عمل Kisay gall y kisay naam nun barbar duhran da amal
ہیر پھیر Her pher	The process of rotation	گردش دا عمل Gardish da amal

ہولی Holi	the festival of Holi	ہندواں دا ایک تہوار Hinduan da ek tehwar
ہنڈولا Handola	song sung on a swing	برسات دا گیت جو جھولے اتے بیٹھ کہ گایا جاندا اے Barsat da geet jo khuly utay beth keh gaya janda ey
ہلا شیری Hala sheri	The act of giving courage, praise	ہمت دلان دا عمل ، شاباش ، آفرین - Himmat dilan da amal, shabash, Afreen

4.5 Cultural Differences

Cultural differences also played a significant role in the variations between both WordNets, contributing to differences in alignment. Many words in Punjabi WordNet were culturally specific to Punjabi culture and lacked direct equivalents in English culture and language. Due to the absence of these words and concepts in English, alignment became challenging. In some cases, the words were missing entirely, or they were translated using phrases, leading to variations in the alignment process.

Connotations also differed based on cultural aspects. A word that has a positive connotation in the English language and culture might carry a negative connotation in Punjabi language and culture. For example, the word "girlfriend" may have a positive connotation in English but not in Punjabi.

Cultural differences were particularly evident in relation to kinship and familial terms. While aligning Punjabi WordNet with English WordNet, differences emerged in the way relation and kinship terms were used across both languages and cultures. For instance, the English word "sister-in-law" is a single word, but in Punjabi, it has more than one equivalent, indicating variations in the kinship terms and lexical items used for relationships in both languages and cultures. Examples of such variations are provided below.

Table 4.9 Differences in Punjabi and English words in Kinship terms

Punjabi Word	English	Punjabi Word	English
بھتیجا، بھانجا Bhanja, Bahteeja	Nephew	بھتیجی، بھانجی Bhateeji, Bhanji	Niece
نند، پابی، درانی، جٹھانی Nand, Paabi, Darani, Jithani	Sister-in-law	دیر، جیٹھ، بہنوئی Der, Jeth, Behnoi	Brother-in-law
سالا، پنویا Sala, Panvaia	Brother-in-law	پرجائی Parjai	Sister-in-law
تایا، چاچا، ماما، ماسڑ Taya, chacha, mama, masarr	Uncle	ماسی، تائی، پھوپھی، چاچی Maasi, Tai, Phuphi, chichi	Aunty

4.6 One-word Translation

During the analysis Punjabi WordNet was aligned with the English WordNet. Princeton English WordNet shows only one word in the search bar. Variations were found in the English translation of Punjabi words. If a Punjabi word has English translation in two words that word was not found in the English WordNet. Examples of such words are maaf karan معاف کرن has the translation to forgive and nyana نیانا has the translation little baby. Princeton English WordNet helps in the search of only one word so this translation was not found in the WordNet.

4.7 Alignment between English and Punjabi WordNet

Alignment between Punjabi and English WordNet was done using the excel sheet. While developing the Punjabi WordNet the data base of synsets and the lexico-semantic relations was developed in the excel sheet and later on developed in the form of utility software. The excel sheet on which data base of Punjabi WordNet was developed, was further extended to align the

English WordNet with the Punjabi WordNet. Synsets of the Punjabi WordNet were aligned with the synsets of the English WordNet in the excel sheet. While aligning the synsets of Punjabi WordNet with the English WordNet the main aim was to match the same sense of Punjabi and English WordNet.

This alignment of Punjabi and English WordNet matches the exact sense among both languages. This data base developed will help in automatic machine translation and bilingual dictionary in which one sense of a word will be translated and aligned with the exact sense of English language. In result of which, bilingual speakers will be able to comprehend and understand both languages more efficiently. Furthermore, this data base can also be used for English language teaching. L1 reference can be used for the better understanding of L2. This data base can provide examples for providing examples from the L1 of the learners. The examples will include translation of a word, along with the gloss and sentence of both languages. So, in this way the resource developed will assist in the process of language teaching.

The linguistic database created as part of this study holds significant potential for advancing the digital resources of the Punjabi language. Punjabi is often considered a less prestigious and less developed language, primarily due to the limited availability of digital resources. The lexical and digital resources generated through this study can contribute to elevating the status of the Punjabi language in the digital realm.

Throughout the course of this study, several digital tools were developed, including a morphological analyzer, a spell checker for Punjabi, a part-of-speech (POS) tagger, and a semantic tagger. Additionally, a Punjabi WordNet was created, encompassing lexico-semantic relations, gloss, and sentences. Finally, a utility tool was developed, facilitating the alignment of English and Punjabi WordNet, ensuring words, gloss, and sentences accurately match the sense of both languages.

These digital resources, spanning from linguistic analysis tools to a comprehensive Punjabi WordNet, offer valuable assets for researchers, educators, and language enthusiasts. They not only contribute to the preservation and advancement of the Punjabi language but also open doors to a wider range of applications, including NLP, machine translation, language teaching, and more.

5. Discussion

During the analysis, several instances emerged where one English word had multiple equivalents in Punjabi. For example, the English word "day" corresponds to Punjabi equivalents like "دن" (din), "دہاڑ" (diharr), and "دہاڑی" (diharri). Similar occurrences were found for nouns like "پیار" (pyar) and "محبت" (muhabbat), which is translated to the English word "love," and for words like "خوبصورتی" (khubsorti) and "حسن" (hussan), both translating to "beauty." Verbs like "سدھن" (sadhan) and "بلان" (bulan) also had a single English equivalent, "call," as did "نسن" (nassan), "پجن" (pajjan), and "دوڑن" (dorran), all translating to "run."

The role of diacritics in Punjabi became evident during the analysis. Words like "لڈو" (ludoo), referring to a game, and "لڈو" (laddu), describing a type of sweet common in the subcontinent, may appear similar in written form. However, diacritical marks differentiate their meanings. For instance, "پجن" (pajjan) signifies "run," while "پجن" (pujjan) conveys "fry." Notably, written Punjabi often omits diacritics, relying on context for meaning.

English words sometimes have multiple Punjabi equivalents due to variations in number, gender, and case. For instance, the English word "naked" has four Punjabi equivalents: "ننگا," "ننگے," "ننگیاں," and "ننگی." These variations stem from differences in number and gender, highlighting the intricacies of morphology.

Numerous Punjabi words lacked direct English equivalents, primarily reflecting local Punjabi culture. Examples include "چکی" (chakki), "بالچھی" (balchi), "سنڈ" (sund), "سرمچو" (surmachu), and "موہڑی" (mohrri). These

words are deeply rooted in Punjabi culture and have no counterparts in English.

Certain Punjabi words, primarily adjectives, were used as proper nouns in everyday Punjabi speech. For instance, "امانت" (amanat) and "مقدس" (muqaddas) functioned as proper nouns. This interchangeability between adjectives and proper nouns added depth to Punjabi language usage.

Equivalents of Punjabi words often differed from their English counterparts in terms of grammatical category. For instance, "میتھوں" (meethon) in Punjabi is an adverb, whereas its English equivalent, "myself," is not.

The second-person pronouns exhibited differences between English and Punjabi. English had a single second-person pronoun, while Punjabi featured two—singular "توں" (tun) and plural "تسی" (tusi). This variation created challenges when developing equivalents.

As the study delved into the intricacies of word equivalences, it shed light on the dynamic relationship between Punjabi and English. While some words found direct counterparts, others revealed the fascinating interplay of linguistic, cultural, and contextual factors that shape language. These findings serve as a testament to the richness and complexity of the Punjabi language, providing valuable insights for both linguistic research and cross-cultural communication.

During the process of finding equivalents for Punjabi words in English, an interesting observation was made regarding the word "جوان" (jawān), which means "young" in Punjabi. This word had two equivalents in English, one as a noun and the other as an adjective. Such instances highlight the complexity of word equivalences between the two languages, where a word in one language may have multiple equivalents in the other, depending on its grammatical function.

Additionally, certain lexical items, which are typically neutral, became gender-specific due to contextual language use. Words like "پانی"

(pani) for water, "روٹی" (roti) for bread, and "کباب" (kabab) for meat, originally neutral, acquired gender specificity through contextual usage.

Regarding the morphology of Punjabi nouns, it was observed that masculine nouns typically end with "ا" (aa) and form plurals by adding "ے" (ay), while feminine nouns end with "ی" (i) and often create plurals with "ان" and "یاں" although exceptions existed.

Furthermore, while aligning Punjabi words with their English equivalents, it became evident that Punjabi words often had gender-based distinctions, leading to multiple English equivalents for a single word. For example, "بلا" (bila) and "بلی" (bili) in Punjabi both translate to "cat" in English. The presence of "ا" indicated masculine or larger objects, while "ی" referred to feminine or smaller objects in Punjabi.

6. Conclusion

Punjabi is considered an underdeveloped language, primarily due to a dearth of digital resources. This study endeavors to play a pivotal role in enhancing the status of the Punjabi language. Through rigorous research and the development of lexical resources via the alignment of Punjabi WordNet with English WordNet, our study aims to make substantial contributions toward advancing the standing of Punjabi in the digital realm. This study's implications span diverse domains, including language teaching, research, machine translation, and language digitization. The morphological analysis conducted within this study significantly contributes to understanding Punjabi's word formation processes. The POS tagger and semantic tagger open new avenues for linguistic exploration in Punjabi. The lexical resource we have developed, complete with English translations, gloss, sentences and lexico-semantic relations presents significant opportunities to enhance language teaching. It can serve as a valuable L1 reference for English language learners. Furthermore, the lexical resource established in this study becomes a valuable asset for future research endeavors, experimental studies, the development of software applications, and a wide array of studies built

upon this rich lexical database. Lastly, the issue of script standardization in Punjabi remains a challenge. The study highlights the variations in script, with one word appearing in multiple orthographic forms. Further research into this area could lead to the development of a unified and standardized script. In conclusion, this research bears significance for both Punjabi and English language teaching. It is poised to benefit Punjabi speakers by elevating the language's digital presence. The digital resources generated in this study promise to be invaluable for future research in the realm of Punjabi linguistics.

References

Akhter, N., Mahmood, M. A., & Nadeem, M. T. (2019). *Desarrollo de nombres en Punjabi y relaciones léxico-semánticas. Dilemas contemporáneos: Educación, Política y Valores*.

Arslan, M. F., Mehmood, M. A., & Hayat, S. (2019). Corpus-based study on vocabulary profile of Shahmukhi Punjabi language. Contemporary dilemmas: Education, Politics and Values.

Arslan, M. F., & Mahmood, M. A. (2021). Highlighting the Sound Shift in Punjabi Language: A Corpus-Based Descriptive Study. In *Linguistic Forum* (Vol. 3, No. 1).

Atserias, J., Climent, S., Farreres, X., Rigau, G., & Rodriguez, H. (2000). Combining multiple methods for the automatic construction of multilingual WordNets. *Amsterdam Studies in the Theory and History of Linguistic Science Series* 4, 327-340.

Balkova, V., Suhonogov, A., & Yablonsky, S. (2008). Some issues in the construction of a Russian wordnet grid. In *Proceedings of the Forth International WordNet Conference*, Szeged, Hungary (Vol. 44).

Bangalore, S., Rambow, O., & Whittaker, S. (2000). *Evaluation metrics for generation. In Proceedings of the first international conference on Natural language generation*-Volume 14 (pp. 1-8). Association for Computational Linguistics.

Bin, W. (2016). Empirical Study on the Computer-aided College English Translation Teaching. *International Journal of Emerging Technologies in Learning, 11*(12).

Bosch, S. E., & Griesel, M. (2017). Strategies for building wordnets for under-resourced languages: The case of African languages. *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies, 38*(1), 1-12.

Chai, J. Y., & Biermann, A. W. (1997). A WordNet based rule generalization engine in meaning extraction system. In *International Symposium on Methodologies for Intelligent Systems* (pp. 529-539). Springer, Berlin, Heidelberg.

Capel, A. (2012). Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, 3, e1.

Diekema, A. R., Oroumchian, F., Sheridan, P., & Liddy, E. D. (1998). *TREC-7 evaluation of conceptual interlingual document retrieval (CINDOR) in English and French.*

Doddington, G. (2002, March). Automatic evaluation of machine translation quality using n gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research* (pp. 138-145). Morgan Kaufmann Publishers Inc.

Farreres, J. (2005). *Automatic Construction of Wide-Coverage Domain-Independent Lexico Conceptual Ontologies* (Doctoral dissertation).

Farreres, X., Rigau, G., & Rodriguez, H. (1998). *Using wordnet for building wordnets.* arXiv preprint cmp-lg/9806016.

Fellbaum, C. (1998). Towards a representation of idioms in WordNet. In *Usage of WordNet in Natural Language Processing Systems.*

Ghai, W., & Singh, N. (2013). Continuous speech recognition for Punjabi language. *International Journal of Computer Applications, 72*(14).

Hamouda, A. (2013). A study of dictionary use by Saudi EFL students at Qassim University. *Study in English Language Teaching, 1*(1), 227-257.

Hu, X., Graesser, A. C., & Tutoring Research Group. (1998). Using WordNet and latent semantic analysis to evaluate the conversational contributions of learners in the tutorial dialog. In *Proceedings of the international conference on computers in education* (Vol. 2, pp. 337-341).

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., & Kanzaki, K. (2008). *Development of the Japanese WordNet.* In LREC.

Nation, P. (1990). *Teaching and Learning Vocabulary.* New York: Heinle and Heinle.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2017). *Natural language processing: State of the art, current trends and challenges.* arXiv preprint arXiv:1708.05148.

Kulkarni, M., Dangarikar, C., Kulkarni, I., Nanda, A., & Bhattacharyya, P. (2010). Introducing sanskrit wordnet. In *Proceedings on the 5th global wordnet conference (GWC 2010)*, Narosa, Mumbai (pp. 287-294).

Lee, C. K., & Lee, G. B. (2000). Automatic WordNet mapping using word sense disambiguation. In *Annual Conference on Human and Language Technology* (pp. 262-268). Human and Language Technology.

Mandala, R., Tokunaga, T., Tanaka, H., Okumura, A., & Satoh, K. (1998). Ad hoc retrieval experiments using wordnet and automatically constructed thesauri. In *TREC* (Vol. 7, pp. 414-419).

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM, 38*(11), pp. 39-41.

Moldovan, D. I., & Mihalcea, R. (2000). Using wordnet and lexical operators to improve internet searches. *IEEE Internet Computing, 4*(1), 34-43.

Nießen, S., Och, F. J., Leusch, G., & Ney, H. (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *LREC.*

Ordan, N., & Wintner, S. (2007). Hebrew WordNet: a test case of aligning lexical databases across languages. *International Journal of Translation, 19*(1), 39-58.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.

Putra, D. D., Arfan, A., & Manurung, R. (2008). Building an Indonesian wordnet. In *Proceedings of the 2nd International MALINDO Workshop* (pp. 12-13).

Rahman, T. (2002). *Language, ideology and power. Language learning among the Muslims of Pakistan and North India.* Karachi: Oxford University Press.

Ramanand, J., Ukey, A., Singh, B. K., & Bhattacharyya, P. (2007). Mapping and Structural Analysis of Multi-lingual Wordnets. *IEEE Data Eng. Bull.*, *30*(1), 30-43.

Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., & Bogaard, T. (2016). Building event-centric knowledge graphs from news. *Journal of Web Semantics*, 37, 132-151.

Shei, C. C., & Pain, H. (2000). An ESL writer's collocational aid. *Computer assisted language learning*, *13*(2), 167-182.

Thoongsup, S., Charoenporn, T., Robkop, K., Sinthurahat, T., Mokarat, C., Sornlertlamvanich, V., & Isahara, H. (2009, August). Thai wordnet construction. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)* (pp. 139-144).

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997, September). *Accelerated DP based search for statistical translation.* In Eurospeech.

Towell, G., & Voorhees, E. M. (1998). Disambiguating highly ambiguous words. *Computational Linguistics*, *24*(1), 125-145.

Vossen P. (ed.). (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers, Dordrecht.

Zafar, A. (2012). *Developing urdu wordnet using the merge approach.*