

Linguistics and Literature Review (LLR)

Volume 7 Issue 2, 2021

Journal DOI: <https://doi.org/10.32350/llr>

Issue DOI: <https://doi.org/10.32350/llr.72>

ISSN(P): 2221-6510 ISSN(E): 2409-109X

Journal Homepage: <http://journals.umt.edu.pk/llr/Home.aspx>

Article: **Development of Saraiki WordNet by Mapping of Word Senses: A Corpus-based Approach**


Author(s): Sarah Gul, Musarrat Azher, Sana Nawaz

Affiliation: University of Sargodha, Sargodha, Pakistan

Article DOI: <https://doi.org/10.32350/llr.72/04>

Article History: Received: January 27, 2021
Revised: November 26, 2021
Accepted: November 29, 2021

Citation: Sarah Gul, Musarrat Azher and Sana Nawaz. (2021). Development of Saraiki WordNet by mapping of word senses: A corpus-based approach, *Linguistics and Literature Review* 7(2): 46–66.

Copyright Information:  This article is open access and is distributed under the terms of [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

Journal QR



Article QR



Indexing



A publication of the
Department of Linguistics and Communications Institute of Liberal Arts
University of Management and Technology Lahore, Pakistan

Development of Saraiki WordNet by Mapping of Word Senses: A Corpus Based Approach

Sarah Gul
Musarrat Azher*
Sana Nawaz
University of Sargodha, Pakistan

ABSTRACT

The main focus of this paper is to develop the Saraiki WordNet. Saraiki is one of the regional languages spoken in Pakistan and has a unique history of its own. Saraiki language has remarkable similarity with two languages i.e. Punjabi and Sindhi. Saraiki has different dialects and they differ according to the region where they are spoken. This paper uses the Urdu WordNet (Zafar, Mahmood, Shams & Hussain, 2014) as the basis for the formation of Saraiki WordNet. Urdu WordNet (Zafar et al., 2014) is created by UET Lahore and is based on Princeton WordNet (Miller, 1990). Development of Saraiki WordNet is very significant with regard to Natural Language Processing (NLP). Dictionaries or *Lugats* and literary sources such as Poetry and Fiction and non- literary sources like Newspaper of Saraiki language are used for the data purposes and the Urdu word senses are mapped to Saraiki word senses. The method used in this study is mapping and expand approach is used in the mapping process. This study will prove significant in creating bilingual dictionaries in future and this work can be used for further advancement in procedure of the development of the bilingual dictionaries.

Keywords: Mapping, Saraiki language, Expand approach, WordNet

Introduction

Saraiki is counted among the widely spoken languages in the Pakistani provinces of Punjab and Khyber Pakhtunkhwa (KPK). It is a sister language of Sindhi and Punjabi languages and is greatly influenced by both of them. The speakers of this language are scattered across different geographical regions of Pakistan. In each area, their Saraiki speech is influenced by local languages. Hence, Saraiki has incorporated different elements of local languages, which has allowed it to evolve into a distinct but related language (Garcia, 2016).

WordNet is a thesaurus and it is very useful for computational purposes. It can be downloaded and used online. WordNet has different versions in many languages. Princeton WordNet (Miller, 1990) is the first WordNet to be developed in this regard. It is an English language WordNet developed by George Armitage Miller. Previously, dictionaries were used for finding meaning by people and they are available to use for humans only. A WordNet contains not just words and meanings but also incorporates their concepts and examples. This makes it more useful than conventional dictionaries as more emphasis is given to computational methods of word databases. It is an intricate system and consists of a database provided with a complete system of documentation and tracking (Miller, 1990).

*Corresponding Author: musarratazher@gmail.com

The WordNet developed by George Armitage Miller consists of strings of words. Each word has multiple synonyms. These synonyms pertain to one sense of the word. The WordNet incorporates all of their related meanings, senses and concepts. There are a total of 118,000 different word forms in it. There are different word senses totalling around 90,000 and pairs included are totalling around 166,000. The amount of polysemous words is 17% and 40% have set of synonyms. Different categories are distinguished in this WordNet based on different criteria. Nouns, verbs and other categories are mentioned. In terms of parsing system, some 300 prepositions and pronouns are important (Miller, [1990](#)).

Inflectional and derivational morphology is also taken into context in this WordNet. Inflectional morphology is a big part of the WordNet system, it provides the option for seeing the other form as well while on the other hand derivational morphological information is also given a distinct position in it. Different semantic relations are also given importance (Miller, [1995](#)). Some of the semantic relations mentioned are synonymy, antonymy, hyponymy, meronymy, toponymy and entailment. All these relations are assigned specific categories and there are almost 116,000 such relations in this WordNet (Miller, [1990](#)).

The systems provided in this WordNet make it possible to search the required item and find the required category. This, in turn, makes it possible to find the exact meronyms and hyponyms of a given word. It helps in retrieving the information easily and keeping it at hand. The issue of polysemy rises when one language is translated into another. Sometimes, there are multiple meanings provided for one word and it causes problems in determining the proper translation for the said word. It is crucial to take context into consideration in order to find the exact meaning. This WordNet needs a lot of development in this regard as it gives multiple meanings without giving proper consideration to the context and it becomes hard to find the relevant meanings of the words. Algorithms are needed to provide the required context. Sense identification is very important in order to find the exact meaning. Proper contextual representation is needed in a WordNet (Miller, [1990](#)).

Lots of methods have been used to counter this issue in computational linguistics. One way is to limit the discourse. Topical context is another way to solve this problem. Sometimes, local context is also used to solve this issue. Still, a proper system is needed to find the correct meanings according to the context, the absence of which is creating a lot of problems for people using this WordNet. Semantic concordance is very important in creating links in the lexicon contained in a corpus. It is a small-scale method to solve this problem and a large-scale solution of this problem is still required (Miller, [1990](#)).

Aims and Objectives

- The aim of the current study is to develop a Saraiki WordNet using Urdu WordNet as its basis.

Research Questions

This research asked the following research questions:

1. What is the process involved in mapping Urdu word senses to Saraiki word senses?
2. How the word senses of the two languages are aligned to help develop Saraiki WordNet?

Literature Review

Saraiki language is part of the vibrant culture of Pakistan which has different colours according to the region. Saraiki language has been spoken in multiple parts of Pakistan whether it is in Northern or Southern parts of the country. Saraiki language has a long history of its origin and dialects.

To some researchers, lexical knowledge base is very useful in resource development of language. An important lexical knowledge base is the WordNet, which is very useful in language processing. There are many ways to extend this resource. WordNet is one of the most important components of lexical knowledge base. It helps in semantic search, text summarization and Word Sense Disambiguation (Fernando & Stevenson, 2012).

Mapping is one of the ways used to enrich the WordNet and also to develop them. The use of automated as well as manual methods is very important regarding the development of the WordNet. This is why, in the development of this WordNet, the use of manual annotation has been very important. These mappings after the development of wordnet are put online for access (Fernando & Stevenson, 2012).

There are many semantic relations in WordNet such as the ones given below.

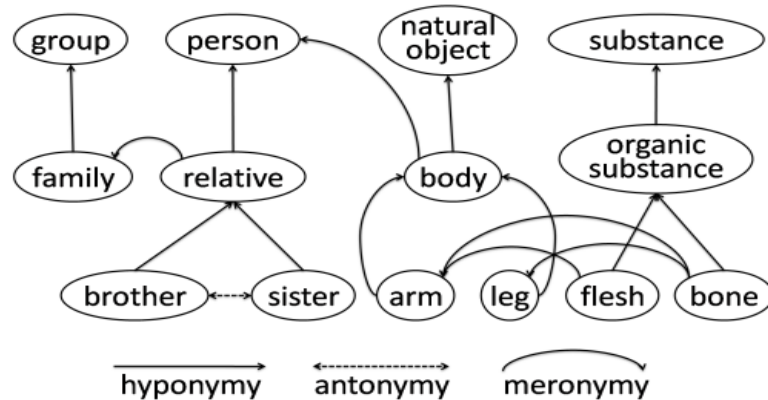


Figure 1.1. Semantic Relations in WordNet. Reprinted from “Nouns in WordNet: A Lexical Inheritance System,” by (Miller, 1998)

WordNet is very important in Natural Language Processing (NLP) and it is an invaluable source in computational linguistics. It works on the basis of a thesaurus. Working to end the problems in everyday dictionaries, Miller (1990) developed the first wordnet. It solved the problems relating to the senses and in the definitions. WordNet consists of lemmas and senses (Artale, Magnini, & Strapparava, 1997).

There are two ways to map senses in WordNet. One is manually and the other is automatically. Automatic method uses the already available resources to construct a WordNet. This method uses Word Sense Disambiguation (WSD), which takes the words collected from bilingual dictionaries and connects them with the WordNet Synsets. There are many dictionaries which use this method to develop WordNet. Many functions related to NLP demand compact ontologies. These ontologies help in information retrieval. Only a few languages have ontologies and many languages still lack in compiling ontologies. It is very difficult to develop ontologies manually as this work demands a lot of time and resources. Researchers use the already available resources to develop the WordNet, as these resources already cover a wide range of lexical knowledge and semantic information. Korean WordNet is among the WordNets developed by this method. It uses the automatic WordNet mapping by utilizing WSD. Korean words from a bilingual dictionary (MRD) are linked with the English WordNet Synsets (Lee, Lee & Yun, 2000).

The mapping process involved in the development of this WordNet used all the heuristics

mentioned above and the decision tree helped in the disambiguation process. All heuristics were used either to link or discard the candidate Synsets. In the instance of Korean WordNet, manual classification was used to link or discard the 3260 candidate Synsets with the senses found in the Korean bilingual dictionary. Precision and coverage were also involved in the process. Precision helps to find the correctly linked senses and coverage indicates the proportion of linked senses (Lee et al., [2000](#)).

There are two methods used to develop a WordNet. One is the ‘expansion’ approach and the other is the ‘merge’ approach. The method employed here is the expansion approach. It has been used previously to develop a number of WordNets. The merge approach is used where extensive resources are available and time constraints are little to nothing.

Several methods are used in the expansion approach to develop a WordNet. These methods include (i) Cross-lingual WSD (ii) Google Similarity Distance (iii) Intersection method (iv) Multiple Heuristic method (v) Combining multiple methods (vi) Assign procedure (vii) Base concepts and (viii) MultiDic tool. Cross-lingual WSD uses both Word Sense Induction (WSI) and Word Sense Disambiguation (WSD). This method is discussed by Apidianaki. The first WordNet developed by using it was the French WordNet. This method creates semantically similar groups and after disambiguation, these groups are placed in their positions in the WordNet. WSI method was used by Apidianaki in the English-Greek corpus. The variations of English words are represented by three Greek equivalents (EQVs) in this WordNet. Every EQV represents a different sense of the given word. To distinguish between each sense and to place semantically similar EQVs in the same cluster, semantic similarity of each pair is calculated (Nadageri & Haribhakta, [2017](#)).

In the above instance of word variation, two words {increase, significant} out of surrounding context features found in cluster 1 representing Greek word διακύμανση with sense fluctuation. Using this approach, Greek equivalents replaced PWN Synsets to create the Greek WordNet. The performance of cross-lingual WSD was found to be quite promising and approximately 72% nouns, 62% verbs, 81% adjectives, and 86% adverbs were correctly distinguished (Nadageri & Haribhakta, [2017](#)).

Another method known as Google Similarity Distance is used to link words with the English Synsets through WSD. To find the suitable link with the Princeton WordNet (Miller, [1990](#)), a similarity was determined between translated Synsets and translated definition in the target language. This method was used in the development of Macedonian WordNet. It was identified that “the result as per the discussion in shows that Google Similarity Distance method has 87% accuracy in assignment of appropriate Synsets. It correctly translates 14,335 English Synsets into Macedonian Synsets” (Nadageri & Haribhakta, [2017](#)).

Intersection method is another method used in the development of WordNet. In this method, synonymy is the main feature as it is responsible for creating equivalence classes. Two WordNets including Macedonian and Romanian WordNets were developed using this method. It involves two rules (Nadageri & Haribhakta, [2017](#)). The first one states that if the original Synset contains at least one monosemous word, then the translation of that monosemous word is sufficient to translate other words in the Synset. The second rule is that if the original Synset contains more than one polysemous word, then the intersection of the translations of each word in the synset forms translation of original synset (Nadageri & Haribhakta, [2017](#)).

Another method is called ‘combining multiple methods’ which combines different ways to develop a WordNet. Homogenous Bilingual (HBil) dictionary is very useful and based on this method. It has word entries in both ways to make it easier to work in both languages. This dictionary helps in linking senses with the WordNet. Other methods have been used also in this way including class method, structural method and conceptual method. Class method uses the processed dictionary and criteria to develop words. The criteria used in this method are the polysemic criterion, hybrid criterion and field criterion. Structural method takes the whole

structure of PW and links it with the Synsets of the Princeton WordNet (Miller, 1990). The criteria involved in this method are the intersection criterion, parent criterion, brother criterion and distant hyperonym criterion. The last method is the Conceptual Distance method. This method deals with the closeness between the meanings of words. It is calculated to find the closeness of words with each other and monolingual dictionary entries are explored. With the accuracy level of 85%, the Synsets are linked to the Princeton WordNet (Nadageri & Haribhakta, 2017).

Spanish WordNet is built using a combination of these methods. The results of these methods are quite encouraging. In Spanish WordNet v.0.0, all Synsets with a Confidence Score (CS) of more than 85% were selected and 10,982 connections were obtained. Combining discarded Synsets having CS less than but near to 85% could be acceptable, as new connections increased the number of connections by 7,244. Finally, Spanish WordNet v. 0.1 with greater accuracy of 86.4% was obtained (Nadageri & Haribhakta, 2017).

Research Methodology and Corpus Development

Expansion approach is the most widely used method in WordNet development. Lexicographers use this method to build a WordNet. This method is one of the ways to connect to another WordNet as well which results in the WordNet carrying the format and properties of the other developed WordNet.

Dictionaries Used in the Current Study

In this study, different sources were utilised. Of these sources, dictionaries played a big part in the development of WordNet. These dictionaries were both monolingual and bilingual. Bilingual Saraiki-English and Saraiki-Urdu dictionaries proved to be very helpful in conducting this study.

Table 3.1. Dictionaries and their Sources

Source	Type	Name of articles or books	Publishers of the books
<i>Books</i>	Dictionaries	Pehli Wadi Saraiki	Saraiki Area Study
		Lugat by Muhammad Saad Ullah Khan	Centre, BZU, Multan.
		Khatran	
		Glossary of the Multani Language by E. O' Brian	Saraiki Adabi Board Multan
		Siraiki English Dictionary by Andrew Jokes	Saraiki Adabi Board Multan
		Dictionary of the Jatki or Western Punjabi language	Religious book and Tract Society Lahore

Urdu WordNet and License

Other resources used in the process included Urdu WordNet Zafar et al. (2014) developed by (CLE) UET Lahore. GCU Faisalabad also bought this resource from (CLE) UET Lahore. For

this study, we acquired this resource from Government College University Faisalabad for purely research purposes. Centre for Language Engineering (CLE) Lahore freely allowed the use of this resource for research purposes.

Sources of Corpora

Corpus was compiled using different sources such as newspapers, stories, essays and poetry. It took a long time to compile this diverse corpus which proved to be very helpful in providing necessary examples and also helped to elaborate the concepts. There are many WordNets which have used the corpus in the process of development of these databases. One such WordNet is the Tatar WordNet Galieva, Nevzorova & Suleymanov (2015), also called the Tat WordNet. It uses the Tatar National Corpus as the source to collect verbs. Due to the ambiguities regarding the semantics of Turkic and Tatar words, there is a need of a comprehensive language source. The Tatar National Corpus helped to find the correct definitions and also helped in creating a hierarchical network in the development of Tat WordNet. Its use spurred the development of the modern WordNet. It also helped to analyze the various syntactic features and hierarchical networks of semantic relations (Galieva et al., 2015).

Developing a WordNet-like thesaurus of Tatar verbs allowed us to combine the experience of the traditional Tatar lexicography and modern information technologies. The Tatar National Corpus played an important role in building the Tatar WordNet. The use of corpus technology enabled us to create a resource that reflected adequately the distribution of Tatar words and their lexical-semantic variants in real contextual environments (Galieva et al., 2015).

The real use of language in corpus is beneficial as it yields adequate data to provide definitions in the WordNet. Next is the development of synsets which requires a lot of data and analysis. The Tatar National Corpus helped to find the correct pairs and synsets. These relations and pairs proved pivotal in the analysis and processing of data. The Tatar National Corpus takes into account the verbs but other parts of speech can be processed as well (Galieva et al., 2015).

Disciplines of Corpora

Different sources were used to collect the data manually. Corpora were developed using either the automatic method or the manual method. In one study by Giampieri (2019), the manual method was found to be far more reliable, although very tiresome in processing. These diverse sources helped to provide the required amount of data needed to establish the proper use of language. The different disciplines used in the study are given below.

Table 3.2. Discipline Wise Division of Corpus

Serial No.	Disciplines used in the study
1	News or Media
2	Fiction
3	Essays

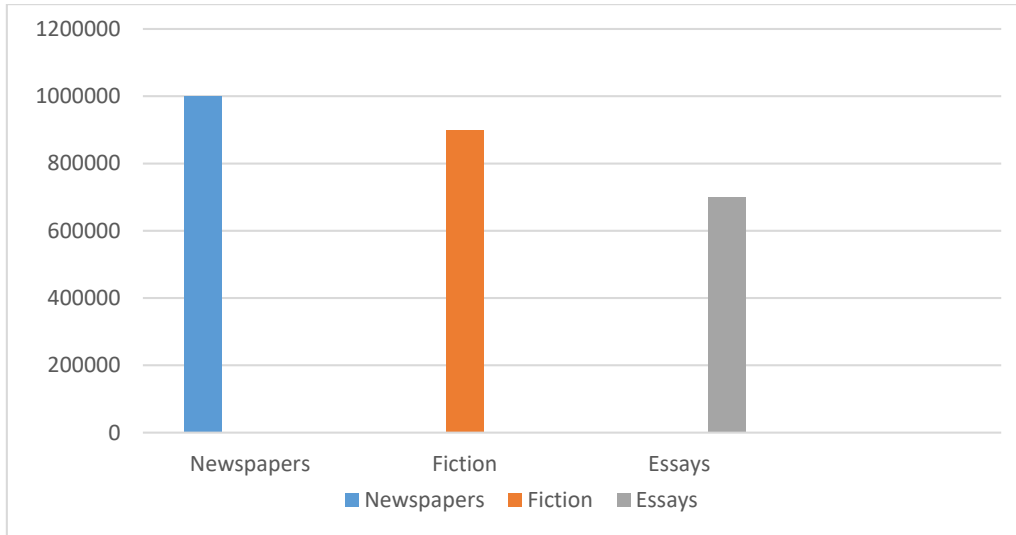


Figure 3.1. Corpora and the Division of the Data

Table 3.3. Types of News Reports and their Number

Newspaper	Type	Number of reports
<i>AFP Newspaper</i>	Sports	200
	International	400
	National	200
	Weather Reports	50

Table 3.4. Number of Sources and their Details

Source	Type	Name of articles or books	Publishers of the books
<i>Wikimedia</i>	Fiction	Wasan Mang Eid da taufa kia behjan Kharay chardi saik by Javed Asif Tez ro di takar by Sufi Faiz Muhammad	Jhok publishers
<i>Books</i>		Dilchasap Lu lu saik sray wich rehn Kalam e Shakir by Shakir Taunsvi Nukar Natak by Khalid Iqbal Maskar by Khalid Iqbal Lai Har by Javed Asif	Saraiki Adabi Board Multan Jhok Printers Multan
<i>Wikimedia</i>	Essays	Saraiki Wikipedia Places Saraiki Wasaib by Zahoor Ahmed Dhareeja Saraiki ty Saraiki Wasaib Saraiki ilmi adbi Khazana Saraiki language Personalities	Saraiki Adabi Board Multan

Table 3.5. Types of Data and the Number of Contents

Corpus	Type	Total content
<i>Saraiki Corpora</i>	Newspapers	1 million
	Fiction	900k
	Essays	700k

All of the above sources were utilized to obtain the required data. These corpora helped in cross-checking the data and in the authentication of the mapping process. Corpora were compiled in Word. This Word Document has untagged corpora and all the data needs to be properly tagged. The untagged data is still unrefined and needs to be properly edited and saved in the Word Document.

Converting Data into Machine Readable Form

All the data gathered from different sources was later converted into machine readable format. For this task, different tools and methods were used. These books were initially in hard form and it took a tremendous effort to convert them to machine readable format.

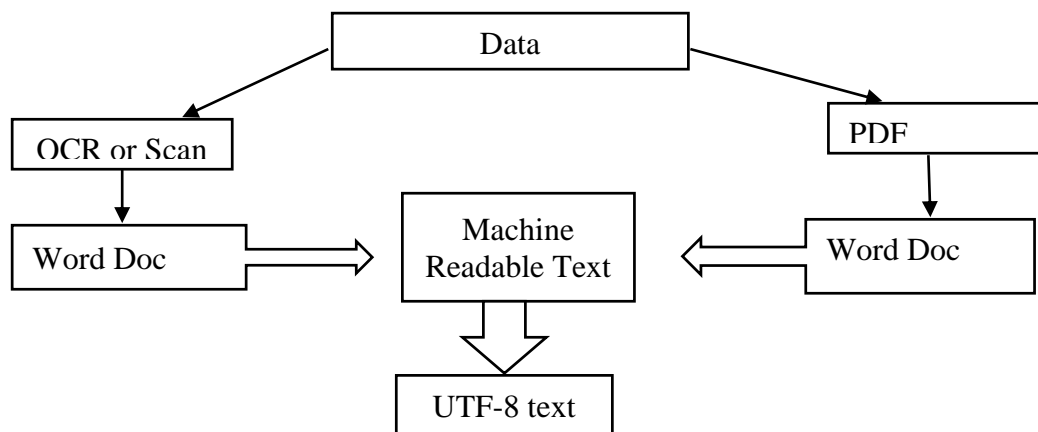


Figure 3.2. Process of converting data into machine readable format

Firstly, all the books in the data were scanned by using the HP DeskJet All-In-One Printer and put through a process of converting them into PDF by using the iLovePDF website. The website converted the data into PDF format. In some instances, OCR was also done by utilizing Google Lens, which helped in segregating the text from the images. Later on, this text was pasted into a word file and a word document was developed using this method. This data was later combined with the data taken from other sources, such as internet.

Coding the Corpus

The corpus of newspaper was given the code NP. The corpus of fiction was coded as FT and the corpus of essays as ES. All these codes were properly mentioned for each of the corpus and during the process of compilation, these codes helped to identify the different sources used in the corpus.

Universal Tag Set

The tagging of data was done using the POS tag set. For this purpose, especially designed POS Tag sets are freely available. These POS tag sets helped in tagging data and categorized them

in proper grammatical categories. POS tag sets can be developed from scratch or they can be downloaded as well. There are many forms of POS tag sets. Some POS tag sets are made for specific reasons and some are made for many languages. The POS tag sets made for many languages can be used for tagging multiple languages and they are known as Universal Tags. The information about the tags of nouns, verbs and other parts of speech helps to know about the grammatical categories of collocated words. For example, knowing nouns from POS tagging helps us to know the adjectives and other grammatical categories. The placement of noun in a phrase tells us about the nature of the phrase. POS tagging helps in many ways. One of the many benefits of POS tagging is that it helps in the process of information extraction about people and organizations, which are all named entities. Another benefit is speech recognition and co-reference resolution (Jurafsky & Martin, [2019](#)).

Benefits of POS Tag Sets

Universal taggers are used for many languages and this is their main benefit. There are many languages which are tagged with universal tag sets. These tag sets have been developed by many researchers. All the languages tagged with universal tag sets create a kind of database where they can be compared and mapped together. Two universal taggers used are Universal Dependencies and Google Universal POS tagger. Both of these POS taggers are easily available and can be used in multiple documents. These are refined POS taggers which provide clarity regarding the use of grammatical categories. Universal Dependencies tag set has 16 tag sets and these can be modified further to add grammatical categories of different languages (Nivre et al, [2016](#) as cited in Jurafsky & Martin, [2019](#)).

Google Universal Tag Set

The POS tagger used in this study was the Google Universal Tagger. It is quite helpful as it gives us basic details regarding POS tagging. The Google Universal tag set consists of twelve POS tags. It not just provides tag sets but also performs the mapping of 25 treebank tag sets from different languages. These mappings prove helpful in providing the tag sets needed to compare different languages. After combining it with the other main tag set, we created a database of almost 22 different languages in the same place. To check the benefits and use of this tag set, it went through many experiments. All the treebanks were checked through the Universal POS tag set to know its authenticity. For unsupervised grammar induction and parser, the Google Universal tag set was utilized (Petrov, Das & McDonald, [2011](#)).

Table 3.6. Google Universal Tag Set

Categories	Types	POS Tag
Verbs	All tenses and modes	VERB
Nouns	Common and Proper	NOUN
Pronouns		PRON
Adjectives		ADJ
Adverbs		ADV
Adpositions	Prepositions and postpositions	ADP
Conjunctions		CONJ
Determiners		DET
Cardinal numbers		NUM
Particles or other function words		PRT
Other	Foreign words, typos, abbreviations	X
Punctuation		. -

The table shows the Google Universal tag set which is complete and contains all basic grammatical categories. This is a basic tag set and it is very helpful in providing mappings between different languages.

Table 3.7. Word Types and Word Tokens in the Corpora

Corpora	Word Types	Word Tokens
Complete	50231	1310260

This table gives us the facts about the total number of words in the corpora. Word type denotes the individual words, while word token denotes the frequency of occurrence of these words in the corpora.

Results and Analysis

Development of the WordNet

The process of the development of the Saraiki WordNet was marked by various issues. WordNet was developed by using the expansion approach. For this purpose, a complete WordNet was needed to help in the mapping process. Urdu WordNet (Zafar et al., 2014) was used in the mapping process. Saraiki words were taken from the corpora developed from news reports, poetry and other sources. Excel sheets in Microsoft Excel were used to store the basic database. These sheets were first loaded with Urdu WordNet (Zafar et al., 2014) acquired from CLE, UET Lahore. This Urdu WordNet (Zafar et al., 2014) was received in UTF-16 format in Notepad, which was later loaded in Excel sheets. Relevant labels were also provided and data was refined to fit according to the requirements. This WordNet was later used as pivot for further work on developing Saraiki WordNet.



Figure 4.1 Raw form of the Urdu WordNet

Urdu WordNet was provided in raw form. It needed some refinement to be loaded into the excel sheets and also needed to be properly labelled. This WordNet has IDs, POS, concepts and examples. Excel sheets were used to organize the data and to link it with the Saraiki data.

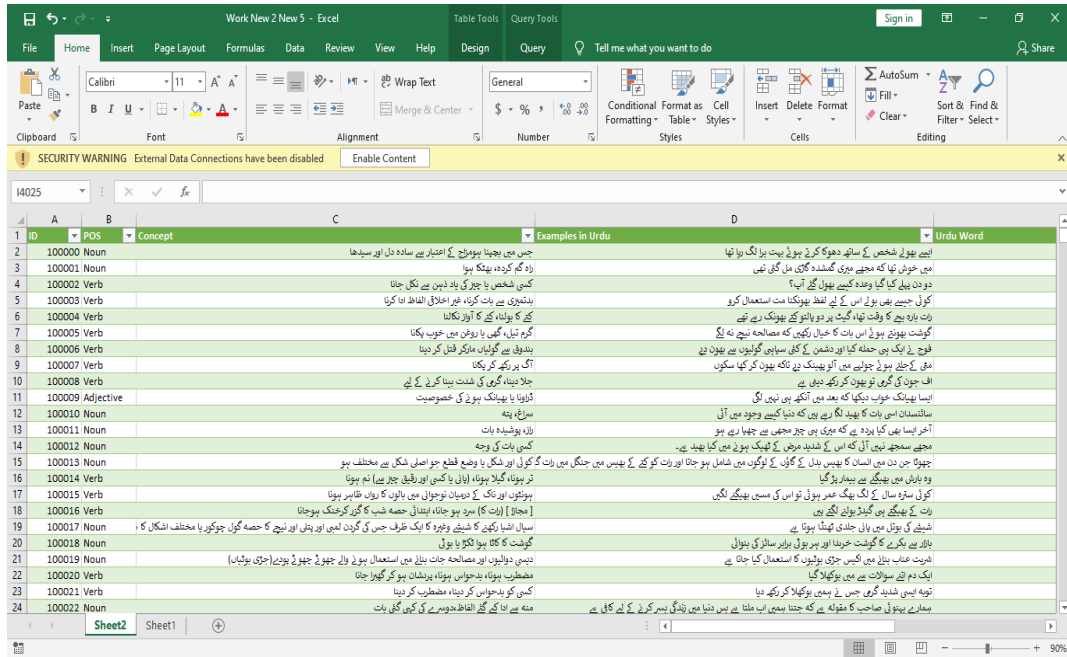


Figure 4.2. Final form after Urdu WordNet is loaded into Excel sheets

Translation of Urdu Entries

After the loading of data, the process of translating Urdu entries into Saraiki began. Saraiki translation of Urdu entries took into account all the senses of the words and no concept was left out. Hence, there was less confusion and retaining the clarity of senses remained the utmost priority at that stage. These translations were made with the help of native speakers and bilingual dictionaries. These sources helped in doing literal translations of Saraiki words and they also helped in other processes. The literal translations were all documented in the Excel sheets. Later on, they were compared with the corpus for another round of determining their authenticity and also to root out any mistakes or false translations.

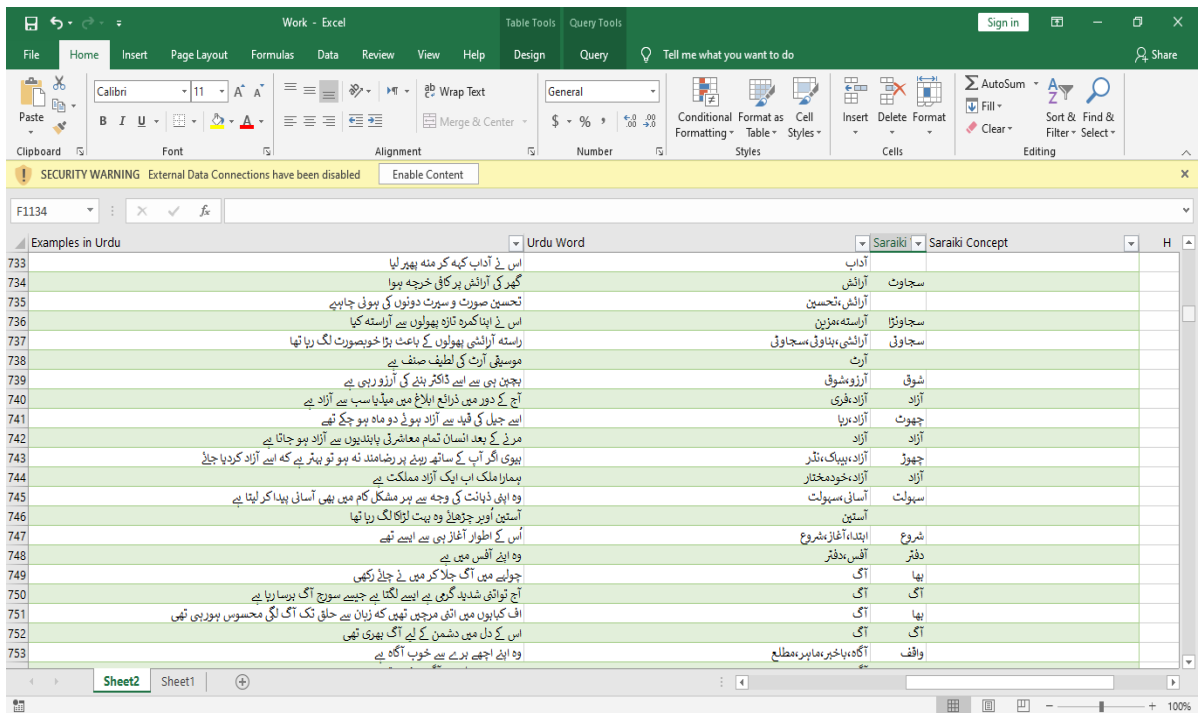


Figure 4.3. Literal translations of Urdu entries

These entries were stored in a separate database and afterwards helped in finding the correct senses. The translations were the starting point in the mapping process and comprehensive translations were very important at this point.

Role of Corpora

The next process involved the preparation of the corpus. The corpora used was of the three kinds already explained in the chapter on corpus compilation. The corpora composed of different sources which proved to be very diverse and helpful in finding the correct usage of words and also in the process of translation. The translations were cross checked from the corpora and incorporated in a different database. These translations were mapped with the Urdu words. In the mapping process, the previous literal translations helped a lot as they provided the corner stone on which we decided the suitability of the sense. Suitable senses were later added to Saraiki words, side by side with Urdu words, as final translations of Urdu words after their evaluation based on literal translations and the corpora.

Process of Encoding

The corpus helped to find the correct literal translation of words and also acted as the backup resource geared to provide relevant examples and concepts. It is a long process to determine the literal translation of words from the corpus. The corpora were compiled first in Microsoft Word, with all the relevant words and information. The data was cleaned and stored in Microsoft Word files for future use. It was in the form of UTF-8 format in Notepad. The data in Notepad helped in AntConc analysis (Anthony, 2019). After loading the data in AntConc software (Anthony, 2019), the corpora appeared in this tool ready for analysis. In the different tabs mentioned for different types of analysis, wordlist is among the most important ones. It provides us with the frequency of the words, that is, how many times a single word appears in the corpus. The resultant list is cloned and used for the purpose of analysis. Moreover, it is also used to find the translations of the words and to cross-check them as well.

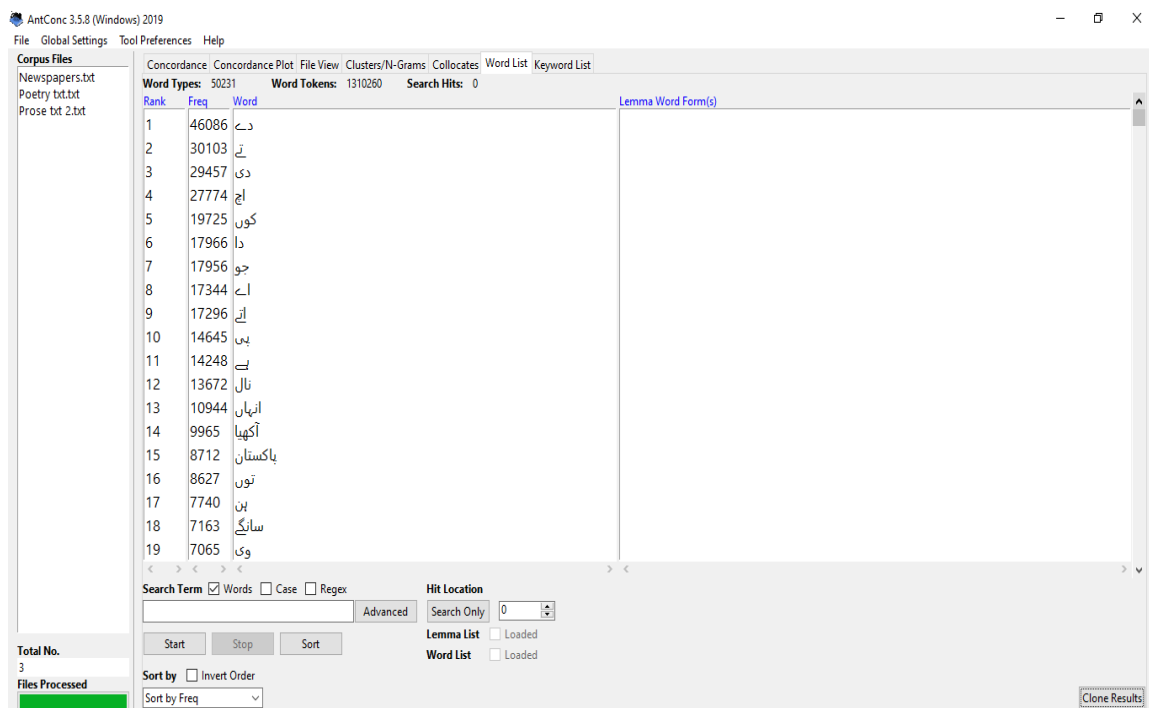


Figure 4.4. Results obtained after loading data into AntConc for analysis

Frequency

Wordlist provides the details about each word used in the corpus including its frequency, which tells us the number of times a word is used in the corpus. Each corpus was loaded into the AntConc software (Anthony, 2019) and its frequency was noted down. The frequencies of all the entries were combined to create a complete list of all the words in the corpora. The wordlist helped to find the correct and reliable senses of Saraiki words. As this wordlist was based on the live use of language, it is a reliable and trustworthy source to be used in the WordNet.

Rank	Freq	Word
1	46086	دے
2	30103	تے
3	29457	دی
4	27774	اچ
5	19725	کوں
6	17966	دا
7	17956	جو
8	17344	اے
9	17296	اے
10	14645	ہی
11	14248	ہے
12	13672	نال
13	10944	ابہاں
14	9965	آکھیا
15	8712	پاکستان
16	8627	توں
17	7740	ہن
18	7163	ساہگے
19	7065	وی
20	6729	وچ
21	6217	اس
22	5997	آباد
23	5770	اسلام
24	5535	وزیر
25	5391	کتبی
26	5251	کتیے
27	5093	پکا

Figure 4.5. Wordlist results

They were later cloned and a list appeared with the ranks, frequencies and words. All this information helped in cataloguing the entire wordlist for later use. The wordlist was used to compare with the literal translation of the words and the words which are used in the corpora are later used in the WordNet.

Urdu Word	Saraiki	Saraiki Concept
اس کے گھر کے صحن میں کئی بونے لگے ہوئے تھے	بولے، پلاٹ، پونا	بولے
انہوں نے گاؤں سے دور ایک نیا پلاٹ تعمیر کروایا	پلاٹ	
وہ پلاٹ میں کھانا ڈال رہی تھی	پلاٹ	پلاٹ
لوہے کی ایک بڑی پلاٹ بنا لی گئی تاکہ نام لکھوایا جاسکے	پلاٹ	پلاٹ
اس کی ظاہری شکل پر مت جاؤ وہ اندر سے بہت ہلید ہے	گندا، ہلید، گندا، جس	گندا
وہ ہم سے ساہلگ میں ہوا پھر رہا تھا	ہمپ	ہمپ
وہ ہم سے پھول پھر رہا تھا	ہمپ	ہمپ
پورے گاؤں میں ہائی کا ایک ہی ہمپ تھا	ہمپ	ہمپ
پوسٹر پر اس کی تصویر نمایاں تھی	اشتراک، ہمپٹ، پوسٹر	اشتراک
دفتر میں اس نے گرگ سے مطبوعہ کاغذات کے ہمپٹ منگوائے	ہمپٹ	ہمپٹ
اس نے ہندردن بعد لاہور آنا ہے	ہندردن	ہندردن
اس شناسائی کے سبب اسے یہ دن دیکھنے پڑے	پہچان، شناسائی، واقفیت	پہچان
بچہ آہستہ آہستہ چیزوں کے درمیان فرق کو پہچانتا ہے	پہچان، پہچان، تمیز، شناخت، فرق	پہچان
اس کے ہاتھ پر چوٹ کا نشان اس کی پہچان بن گیا	پہچان، نشان	پہچان
ڈاکٹروں نے مرض کی تشخیص کر لی	پہچان، تشخیص، شناخت	پہچان
مسلمان کی پہچان ہے کہ وہ جیوت نہیں بولتا	پہچان، شناخت	پہچان
پہلی اولاد اور پہلی کتاب کا کوئی بدل نہیں	پہلا، پوسٹ	پہلا
اچھے وقتوں میں لوگ بھی بیلے ہوتے تھے اور زمانہ بھی اچھا تھا	گلا، پہچان، پونا	گلا
وہ ہائی میں پہلی پوسٹ پر فائز ہے	پہلا، فائز	پہلا
اس کا گھر مسجد کے دائیں جانب ہے	پہلو، جانب، رخ، طرف	طرف
گنبد کے پر پہلو پر خطاطی کی گئی تھی	پہلو	پہلو

Figure 4.6. Saraiki translations

The translations were added to the database after thorough evaluation. These translations were added under the label of Saraiki words.

Concepts

The next step was the addition of concepts which provided the unique explanations of these words. Concepts were provided with the help of the corpora and native speakers. The explanations helped to provide the bases for the addition of further examples. They furnished the meanings and explanations of the literal senses already provided in the WordNet. These explanations were clear and provided complete meanings of these words. Concepts were not ambiguous and solved the problem of the similarity of senses. Whenever a new WordNet is compiled, there is always an issue of similar senses of words. So, in order to solve this issue, corpora were consulted to provide the clarity of meaning and also to remove the ambiguities. The compiled corpora helped not only in providing clarity but also in providing related examples from the live use of language.

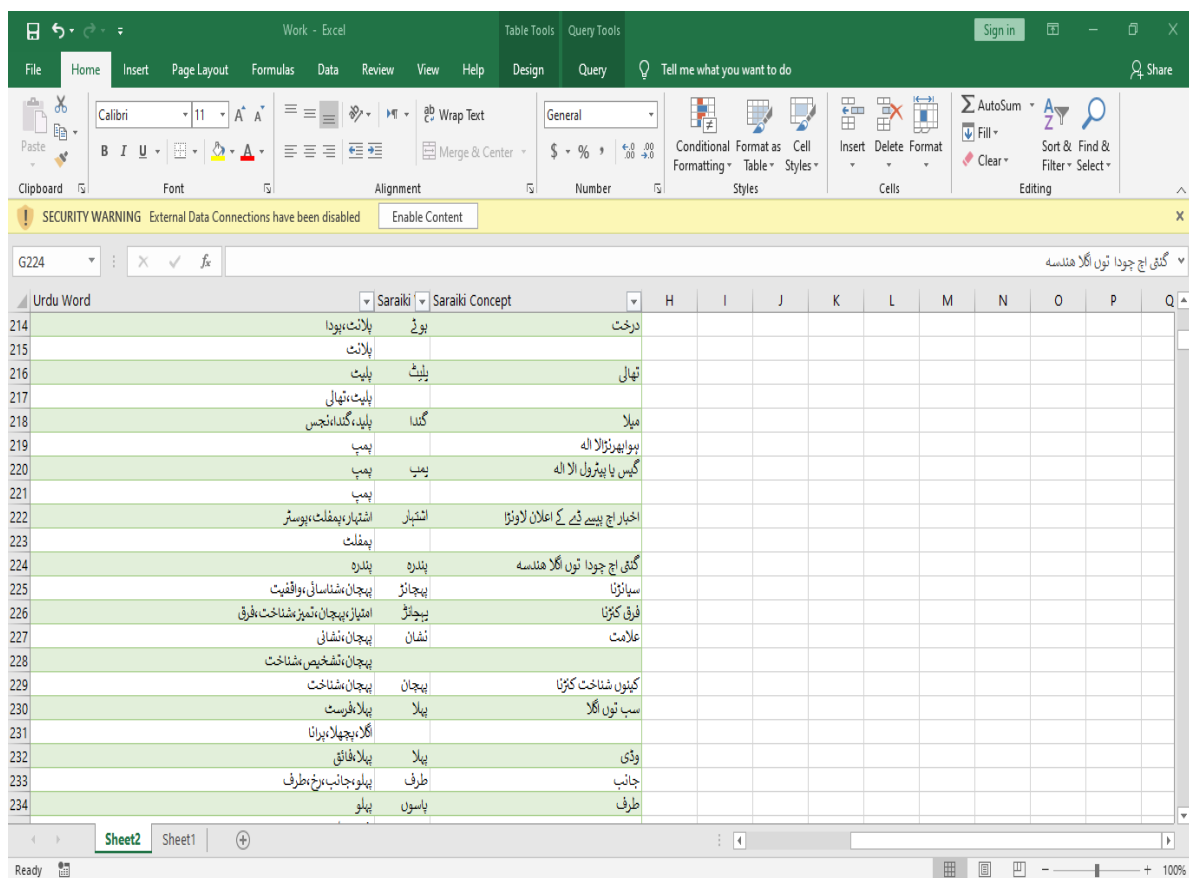


Figure 4.7. Use of concepts to give additional information

The concepts helped to end the ambiguities. They were added right beside the Saraiki senses and together they helped to provide the basic explanations of the Saraiki words.

POS Tags

The next important step was providing POS to the senses of words. After loading the corpora into AntConc (Anthony, 2019), frequency was generated in the wordlist and it revealed the exact number of times a word appeared in the said corpora. The results from the wordlist were later cloned and sent to the Microsoft Word file. The file was used as the basis for the tagging

process, which involved an Urdu Tags set that helped in tagging the relevant and suitable part of speech as the grammatical tag to Saraiki senses. The words in the corpora were tagged according to their grammatical category. Later on, these tags were used to provide Saraiki senses from the corpora with POS categories of verb, noun or adjective.

Urdu Word	Saraiki	Saraiki Concept	POS Tag	I	J	K	L	M	N	O	P	Q
214	پلاٹ، پودا	بوٹے	دخت Noun									
215	پلاٹ											
216	پلیٹ	پلیٹ	نہال Noun									
217	پلیٹ، نہال											
218	پلیٹ، گندا، گچس	گندا	میرا Adjective									
219	پمپ		پوہیزنالا الہ									
220	پمپ	پمپ	گیس یا پیتروں والا الہ									
221	پمپ											
222	اشتبہار، ہمفلٹ، پوسٹر	اشتبہار	اخبار جیسے ڈے کے اعلان لائونڈا									
223	ہمفلٹ											
224	پندرہ	پندرہ	گنتی جی چوڈا توں آگلا ہندسہ									
225	پہچان، شناسائی، واقفیت	پہچانز	سیانزا Noun									
226	پہچان، تہیز، شناخت، فرق	پہچانز	فرق کڑنا Noun									
227	پہچان، نشان	نشان	علامت Noun									
228	پہچان، تشخیص، شناخت											
229	پہچان، شناخت	پہچان	کینوں شناخت کڑنا Noun									
230	پہلا، ٹروسٹ	پہلا	سب توں آگلا Adjective									
231	آگلا، پہچان، پرانا											
232	پہلا، خاکی	پہلا	وڈی Adjective									
233	پہلو، جانب، رخ، طرف	طرف	جانب Noun									
234	پہلو	پاسوں	طرف Noun									

Figure 4.8. POS is provided right beside the concepts

The process of tagging provided us with correct grammatical categories which, in turn, helped us in the mapping process. The correct grammatical categories were very carefully mapped with the Urdu WordNet (Zafar et al., 2014).

Examples

Relevant and suitable examples were assigned to the senses. These examples were taken from the corpus and they provided us with the context exemplifying the use of language. For tracing examples, concordance lines were used to find the relevant word and then the whole sentence was incorporated into the WordNet. Concordance in the AntConc software (Anthony, 2019) helped to find out all the relevant queries related to a word. With context in sight, it becomes far easier to find the suitable query which has all the qualities of a word and does not leave out any meaning.

After loading the data in AntConc (Anthony, 2019), the wordlist was created. The results were cloned and the words chosen were later processed through the concordance procedure. This procedure occurred in the Concordance tab of the tool. The concordance process helped to find the context of the words and the most suitable sentence was chosen which best explained the given word. Full sentences were recorded in file view and added to the WordNet.

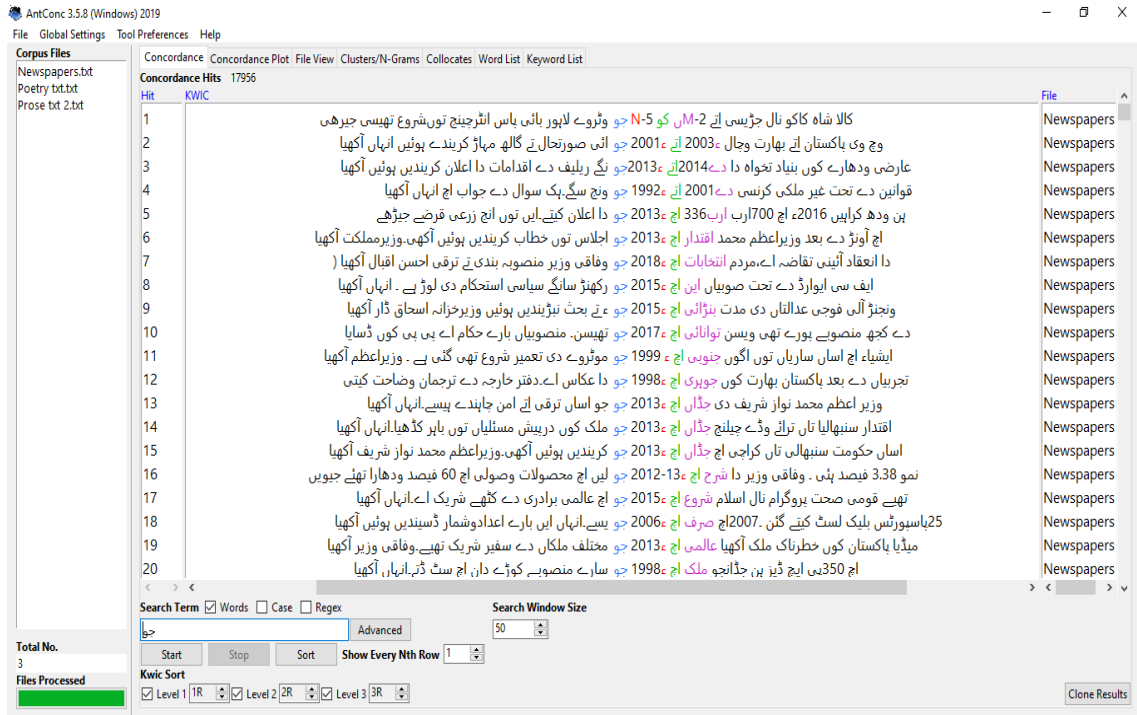


Figure 4.9. Results of concordance

Concordance resulted in seeking different instances of the word used in various contexts. The context which matched the sense of the word in Saraiki language was chosen for the WordNet. Concordance line is the parameter to find the suitable word to add in the WordNet. The process incorporated all three corpus files. It helped in searching the whole corpora and did not leave out any file.

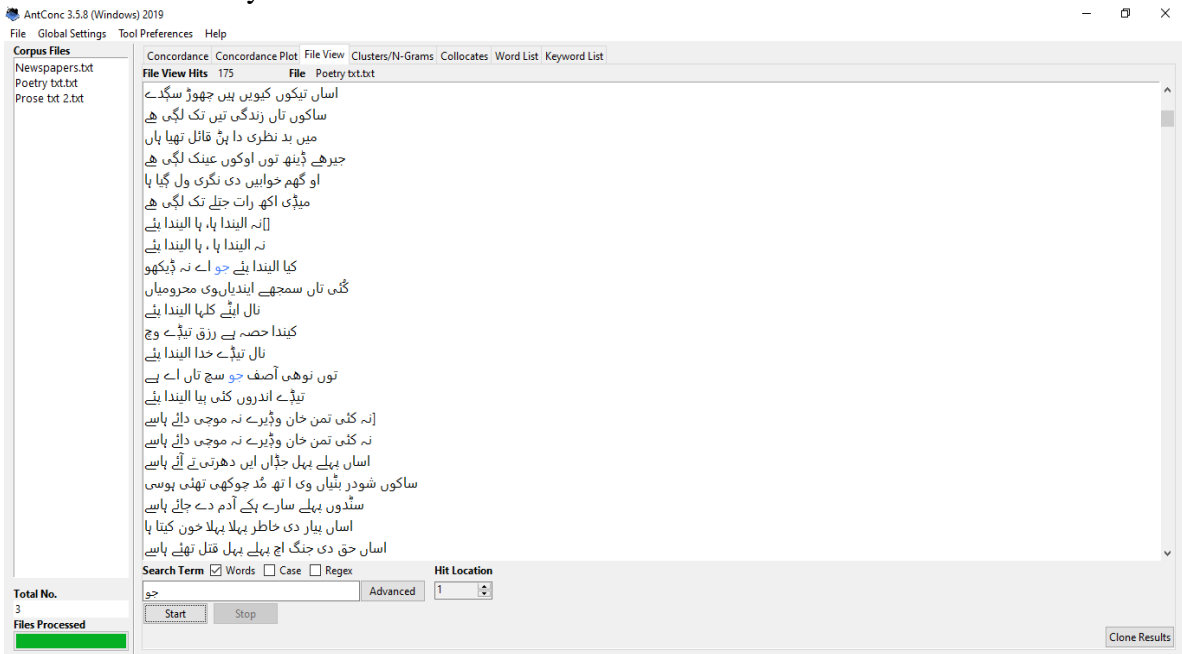


Figure 4.10. File view of the corpora

File view provides the details regarding the use of words in the corpora. When a certain word was chosen from the wordlist to find its concordance, we searched the related words to find the different instances of that word. By clicking at one, the use of that word in different sentences of the corpora was determined. The one instance which was most suitable to the sense was added to the developing WordNet.

Mapping of Senses

WordNet senses of both Saraiki and Urdu words were mapped. The mapping process involved both the Urdu and Saraiki words in order to determine which of these words were mapped and which of them were not. Urdu senses from Urdu WordNet (Zafar et al., 2014) were already loaded into the Excel sheets. Saraiki words, which were chosen from the corpora, were added to the data base. Both Urdu and Saraiki words were later compared. Some of the senses matched, while some did not. There should be clarity in the meaning and translation of the words so that no mistakes are made. Only suitable words were aligned with Urdu words. The mapping process manifested the similarity and differences between the two languages. There should be similarity with regard to the use of language in word senses as well as in other categories. The mapping of senses gave us information regarding the overall similarity between different words based on their literal translations. The senses were all based on Saraiki literal translations and the corpora. Some of the senses were successfully mapped, while some were left out. Successful mapping resulted in the mapping of senses where words were all properly aligned based on their senses, concepts and examples. Any discrepancy between these categories resulted in no match result. Some of the ambiguous and confusing entries were removed from the developed WordNet. The entries which were most compatible with the Urdu word senses were chosen and the rest were removed to avoid any problem. Table 4.1 shows the word senses found in the Urdu WordNet (Zafar et al., 2014).

Table 4.1. Urdu WordNet and its Senses

Urdu WordNet	Number of Word Senses
Noun	2696
Verb	1271
Adverb	97
Adjective	1067

The unmapped word senses were considered no match and it is mentioned as such in the WordNet. There were some words which were hard to map. These words were aligned side by side while keeping in view each and every concept and grammatical category. If the translation of the words were not found in the corpora, they were removed from the WordNet and were not aligned with any source. The words which are mapped are all shown in the Saraiki word part and they are later mapped with the Urdu WordNet (Zafar et al., 2014). It is important for the authenticity of the data that information is taken from the corpus. Without corpora, it would be hard to find the instances of the use of Saraiki words.

Table 4.2. Some Examples of Mapped Senses

ID	POS	Concept	Examples in Urdu	Urdu Word	Saraiki Word	Saraiki Concept	POS Tag	Examples
100011	Noun	پوشیدہ بات	آخر ایسا بھی کیا پردہ ہے کہ میری ہی چیز مجھی سے چھپا راز، رہے ہو	بھید، پردہ، راز	راز	چھپی ہوئی گل	Noun	ڈان لیکس دے معاملے اُتے راز دی گالھ لکائی کاننی گئی

ID	POS	Concept	Examples in Urdu	Urdu Word	Saraiki Word	Saraiki Concept	POS Tag	Examples
100017	Noun	سیال اشیا رکھنے کا شیشے وغیرہ کا ایک ظرف جس کی گردن لمبی اور پتلی اور نیچے کا حصہ گول چوکور یا مختلف اشکال کا ہوتا ہے	شیشے کی بوتل میں پانی جلدی ٹھنڈا ہوتا ہے انسان کے منہ سے نکلے دو میٹھے بول کسی کے دکھ کا گفتگو کے مداوا کر سکتے لیے بولنے کا عمل	بوتل، صراحی	بوتل	کے شے نون بند کرن الی	Noun	سی پیک نال پاکستان تو انائی دے جن کون بوتل اچ بند کرنڑ اچ کامیاب تھی گیا ہے
100024	Noun	سمعی کسی کے دکھ کا گفتگو کے لیے بولنے کا عمل	جتنی جلدی ہوسکے دو بول پڑھوا کر بیٹی رخصت کرو	بول، قول	لفظ	منہ چوں نکلی گل	Noun	آئین دے پک پک لفظ دا تحفظ اتے احترام یقینی بنڑایا ونجے غیر شاتشہ بولی بولنڑ آلا کڈاپیں پاکستانی عوام دا لیڈر نی تھی سگدا میں اللہ کون حاضر ناظر جانڑ تے اکھینداں جو جے آئی ٹی دی کارروائی دے حوالے نال کہیں وی ادارے دے کم اچ رکاوٹ پاتی نہ ہی اگی تے پیساں
100025	Noun	صیغہ نکاح، نکاح پڑھوانا	جتنی جلدی ہوسکے دو بول پڑھوا کر بیٹی رخصت کرو	بول	بول		Noun	
100028	Verb	موسوم کرنا، نام رکھنا وغیرہ	ہم جھوٹے شخص کو دروغ گو بولتے ہیں	بولنا	اکھینداں	کسے نون نا ڈیونا	Verb	

The total number of words which were mapped to the Urdu WordNet (Zafar et al., 2014) was also counted. All parts of speech were counted in the process. Urdu WordNet (Zafar et al., 2014) was first loaded in the Excel sheets and later Saraiki word senses were also loaded into the sheets. The remaining data of POS tags and examples was also added into the developed WordNet.

Table 4.3. Number of Mapped and Unmapped Senses

<i>Urdu WordNet</i>	5132
<i>Saraiki Mapped Senses</i>	2910

Unmapped Senses

There were a number of senses which remained unmapped. These unmapped word senses were removed from the WordNet. Not many senses of Saraiki words were found in the Urdu WordNet (Zafar et al., 2014). Occasionally, some word senses from the Urdu WordNet (Zafar et al., 2014) were not present in the literal translation of Saraiki words. Urdu WordNet (Zafar et al., 2014) has a total number of 5132 senses in it and 2910 senses were matched. The remaining Urdu word senses were not present in the Saraiki corpora and were removed from the WordNet.

Conclusion

WordNet is a great source of lexical information. Different WordNets have been developed in the past.

In the current study, Saraiki WordNet was developed by mapping Urdu and Saraiki word senses. Urdu word senses are part of the Urdu WordNet (Zafar et al., 2014) and these were mapped onto the Saraiki word senses. There are many methods used to develop a WordNet. This Saraiki WordNet was developed using the expansion approach, which is one of the best ways to build a WordNet. The expansion approach helps in linking two WordNets and also takes into consideration already existing resources.

This study was limited due to the constraints of the data. Only corpus data was utilized during the process of WordNet development. Corpora exemplifies a limited use of language. Hence, data was limited to the number of instances provided in the corpora.

This study will prove useful for many researchers working on the Saraiki language. Future researchers working on different reports or dissertations can get useful data from this study and its results.

The biggest advantage that this study confers is the corpora developed for this study. The corpora of Saraiki language manifests its use in different settings and in different varieties. The corpora will prove useful for students and researchers working on this language because of its diversity. Any part of the corpora can be used in future researches dealing with corpus analysis or NLP.

This corpus is helpful for building future bilingual dictionaries as well. Bilingual dictionaries contain information related to two languages. This research is based on mapping both the Saraiki and Urdu word senses and paves the way for the future bilingual Saraiki-Urdu dictionary.

References

- Anthony, L. 2019. *AntConc (Version 3.5. 8)[Computer Software]*. Waseda University.
- Artale, A., Magnini, B., & Strapparava, C. 1997, September. WordNet for Italian and its use for lexical discrimination. In *Congress of the Italian Association for Artificial Intelligence* (pp. 346-356). Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/3-540-63576-9_121

- Fernando, S., & Stevenson, M. 2012. Mapping WordNet synsets to Wikipedia articles. In *LREC* (pp. 590-596).
- Galieva, A., & Nevzorova, O., & Suleymanov, D. 2015. Corpus based tatar lexicography: Verbs in tatwordnet. *Procedia - Social and Behavioral Sciences*, 198: 132–139. <https://doi.org/10.1016/j.sbspro.2015.07.429>
- Garcia, M. I. M. 2016. Saraiki: language or dialect? *Eurasian Journal of Humanities*, 1(2): 40-53.
- Giampieri, P. 2019. Manual and automatic corpus compilation: A case study for legal translations. *International Journal of Language Studies*, 13(3): 1-16.
- Jurafsky, D., & Martin, H. J. 2019. *Speech and Language Processing* (3rd ed.). Stanford.
- Lee, C., Lee, G., & Yun, S. J. 2000. Automatic WordNet mapping using word sense disambiguation. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*, 13 (pp.142–147). Association for Computational Linguistics.
- Miller, G. A., (Ed.). 1990. WordNet: An on-line lexical database. *International Journal of Lexicography* 13(3): 235-312.
- Miller, G. A. 1998. *WordNet: An electronic lexical database*. MIT press.
- Nadageri, Y., & Haribhakta, V. 2017. Building WordNet: A Survey. *International Journal of Engineering Science and Innovative Technology (IJESIT)*, 6(4): 17-27.
- Nivre, J., De Marneffe, M. C., Ginter, F., et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1659-1666).
- Petrov, S., Das, D., & McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.
- Zafar, A., Mahmood, A., Shams, S., & Hussain, S. (2014). *Structural analysis of linking Urdu WordNet to PWN 2.1*. Retrieved from <http://www.cle.org.pk/Publication/papers/2014/Structural%20Analysis%20of%20Linking%20Urdu%20WordNet%20to%20PWN%202.1.pdf>