Urdu Conjunct Predicates (N+V) Inventory from Urdu Universal Dependency Corpus

Farhat Abdullah
Lecturer, University of Central Punjab, Lahore

Tafseer Ahmed
Associate Professor, Muhammad Ali Jinnah University, Karachi

Uzma Anjum
Assistant Professor, Air University, Islamabad

Abstract

This research study aims to develop a semantic inventory of Urdu nouns which may serve as a useful resource for developing natural language processing tools. It is an effort towards improving the severely under-resourced status of Urdu. Conjunct predicate is a type of complex predicate where a noun is followed by a light verb and both work as a single syntactic constituent. Conjunct predicate N+V collocation is extracted from universal dependency annotated Urdu corpus i.e., URDU_UD_UTB (Bhat et al., 2017). Resultant data provided adequate information to categorize the pattern of nouns compatible with light verbs in their all-possible morphological forms. This research yields a sizeable repository of Urdu conjunct predicate along with figuring out a range of case markers licensed by N+V collocation as a constituent which does further implication on the volitionality. Resultant mined data can be used in some future research work to train the data in some cross-linguistic computational programs.

Keywords: conjunct predicate, inventory, semantic role, syntactic context, Urdu corpus, Urdu nouns, universal dependency

1. Introduction

This work aims to formulate the inventory of N+V collocation in Urdu. It sees the compatibility of different noun classes with varying morphological forms of light verb lemmas. N+V combination is highly productive in case of Urdu language and could not be documented to date due to the diverse, individual, innovative and creative use of language choice. Natural language processing needs explicit record of single semantic constituents in a language to develop computational tools. Lack of adequate lexical resources poses difficulty in building up useful computational programs which can be facilitated by building a systematic semantic standard for identifying conjunct predicates. These multiword conjunct predicate with semantic and syntactic value can be inserted in lexical resources such as WordNet as a single entity where its need has already been realized. Review of previous studies on complex predicate provided a sound foundation to move forward in a comparatively less explored dimension (Ahmed & Butt, 2011; Butt, 1993, 1995, 2010, Butt & Geuder, 1998; Butt & Ramchand, 2005; Ehsan & Butt, 2020; Kiani, 2013; Mohanan, 1994, 1997).

Urdu possesses flexible word order, like most languages of Indo-Aryan languages, which mostly displays SOV pattern. It is categorized as one of the 20 most spoken languages of world (Eberhard et al., 2020). It is scarcely resourced to develop required computational programs to match the rapid pace of digitalization of other language being used in the same speech community such as English. It results in un-parallel digital growth of two languages in this diglossic situation. This research may be a step towards empowering Urdu digitally and making it technically compatible with English which is used as a second language in Pakistan (Warsi, 2004). Using the semantic inventory of noun in Urdu conjunct predicate and mapping it onto the English WordNet may facilitate the cross linguistic computational processes such as machine learning, automatic learning and different artificial intelligence program where languages are involved.

The study followed Levin (1993)'s classic theory of verb classification. It is guided by the notion that behavior of verb regarding its expression and realization of arguments is mainly dependent on its semantic sense. She based her research work on the notion that behavior of the verb can be employed to understand the aspects of its meaning. This research paved the way towards the evolution of theory framework for lexical knowledge.

Due to the vast range of nouns which collocate with different light verbs, all possible combinations have not been documented to date. Present research aims to result in the formulation of a lexical resource for Urdu nouns which can serve the purpose of a tool in the development of natural language processing programs for Urdu. Unique, highly productive and sometimes unprecedented instances of conjunct predicate in Urdu language drew attention of many researchers who used different paradigms to study the intervening syntactic and semantic factors, but semantic classes of nouns in conjunct predicate have not been fully explored (Butt, 1995; Kiani, 2013; Mohanan, 1997).

The problem with complex predicate is that their all-possible instances were not documented due to their prolific productive nature. Complex predicate which is a distinct linguistic feature of Indo-Aryan languages including Urdu has been explored from different linguistic perspectives to elaborate on its linguistic uniqueness. Present work aims to explore a new dimension of complex predicate (CP). Noun followed by a light verb forms a CP category, which is termed as conjunct predicate (N+V), has not been fully explored to date. This work aims to develop a systematic semantic inventory of nouns which are syntactically and semantically compatible with light verbs in N+V collocation. Development of an exhaustive lexical tool will bridge the knowledge gap which has been a bottleneck to develop natural language processing programs for Urdu.

2. Literature Review

Urdu is an Indo-Aryan language which has a large speech community in the world with a selected status of National language in Pakistan. In the last three decades, even though the need to technically empower Urdu language has already been felt, it is still an under resourced language. Lack of relevant lexical resources has been a hindrance in developing natural language processing tools for Urdu language.

This study does not claim to fight all syntactic and semantic limitations to ensure the perfect semantic inventory of nouns in Urdu. Conjunct Predicate is an intricate phenomenon and sometimes it fails to pass most of the available constituency tests (Carnie, 2012). First, a sizable, annotated Urdu corpus was required to elicit the possible N+V collocation to study the semantic and syntactic compatibilities of nouns with different light verbs. For this purpose, it needed multilayered annotated Urdu corpus so that conjunct predicate can easily be mined. The researchers have chosen UD_URDU_UTB which is universal dependency Urdu corpus (Bhat et al., 2017). The selection of this Universal Dependency (UD) corpus for Urdu may benefit the study in terms of making it more compatible for future computational work on Urdu because Universal Dependency annotated linguistic resources have become a preference of computational linguists in the recent decade for the advanced natural language processing programs. Furthermore, in order to incorporate a local dialect, other Urdu corpora are also investigated to elicit the conjunct predicate (N+V) pattern.

Known with different names, complex predicate, conjunct predicate and sometimes compound verb is a unique phenomenon common to most Indo-Aryan languages such as Urdu, Hindi and Punjabi. This exclusive linguistic feature provoked researchers to explore it from different linguistic perspectives with diverse approaches. Among the very sound works on complex predicates, Hook (1978), Mohanan (1994; 1993) and Butt (1993, 1993a; 1993b) can be presented as very comprehensive and explanative ventures which were further explored in several studies (Butt, 2010; Butt & Geuder, 1998; Butt & Ramchand, 2005). This section aims to finally focus on conjunct predicate which is noun + light verb collocation after reviewing the fundamental ontology of complex predicate along with different epistemological interpretation of the said knowledge.

2.1 Complex Predicate: A Feature of Indo-Aryan Languages

Complex predicates are treated as single syntactic entity, but it comprises of two or more semantic heads. Complex predicates are abundantly found in South Asian languages except Shina language (Alsina ,1993 & Hook, 1974). According to Butt (1995), complex predicate is polyclausal argument structure which corresponds to single functional structure.

Khailna (play), and *Chamkna* (shine) are simple verbs in Urdu. *Maar Daalna* (beat) and *Shuroo Karna* (begin), *Kush karna* (please) are V+V, N+V and Adj+V constructions respectively which are different forms of complex predicate.

N+V and Adj+V are categorized as conjunct predicate, whereas V_1+V_2 is called as compound predicate.

1. *Usne chor ko* ***maar daala***
   S/he-erg theif-dat kill pour
   "S/he killed the thief."
2. *Usne ganaa* ***shuroo kia***
   S/he-erg song begin did
   "S/he started to sing."
3. *Usne apnay maan ko* ***kush kia***
   S/he-erg his/her mother-dat please did
   "S/he pleased his/her mother."

Example number 1 is V_1+V_2 compound predicate where V_1 is the main verb and V_2 is the light verb. Examples 2 and 3 are categorized as conjunct predicates where noun and adjective are followed by a light verb. Butt (2010) is of view that these nouns, adjectives or main verbs are the main predicational element of a complex predicate whereas the light verbs are usually the syntactic head. This light verb does not carry its distinctive or strong semantic domain but interacts with the main verb to convey the complete sense. She added that light verbs do not always form a single syntactic category, but these can easily be distinguished from auxiliaries and polar verbs.

Complex predicate consists of a sequence of predicates which formulates single predication. Constituents of complex predicate share the same tense, aspect and mood as they form single sequential entity without any gap between them. Multiple functions are performed by complex predicate and their semantic value depends on the predicate (constituent) classes, sentence structure and other contextual assumptions. Complex predicate in oceanic languages are explored as two types – nuclear juncture and core juncture – which are further categorized into symmetrical and asymmetrical complex predicates (Aikhenvald, 2006). According to this classification, core juncture symmetrical complex predicate carries out the sequential purpose actions, whereas other specialized form of complex predicate i.e., nuclear asymmetrical gives information related to adverbs of manner.

Two methods of complex predicate formation have been investigated: merger and coindexation (Baker et al., 2002). Merger complex predicates are formed by merging the same kind of lexical and semantic predicates, whereas, coindexation produces a different variety of complex predicate which are difficult to be expressed by simple predicates.

The phenomenon of complex predicate in Bardi, a Nyulnyulan language spoken in North Australia, has also been studied to explicit that there are three types: raising verbs, restructuring predicates and light verbs (Bowern, 2008).

In Hindi, the concepts of standard aspectual complex predicate and reverse aspectual complex predicate have been distinguished by highlighting the position of light verbs. For instance, the following examples have been used by Poornima and Koenig (2009) to elaborate this phenomenon respectively:

4. *Ram-ne Leela-ko tamaachaa maar di-yaa*
   Ram -ERG Leela-DAT slap.M.SG hit give.M.Sg.PF
   "Ram slapped Leela (hit Leela with a slap)."

5. *Ram-ne Leela-ko tamaachaa de maar-aa*
   Ram-ERG Leela-DAT slap.M.Sg give hit.M.Sg.PF
   "Ram slapped Leela (hit Leela with a slap)."

In example (4) which is standard aspectual CP, light verb is the head, but in example (5) which is reverse aspectual CP, main verb is the head of construction. Bukhari (2009) expressed CP as multi-headed which usually consists of more than a single grammatical element where

each part of these elements carries information related to the head. Like many other researchers who looked for complex predicate in different languages, he was also of the same opinion that it is a frequent phenomenon in Indo-Aryan languages. He categorized three structures of complex predicate in Gojri language which is spoken in Azad Jammu and Kashmir, Pakistan. His claim that light verb can also precede the main verb coincides with the research findings of Poornima and Koenig (2009) who proved it in the case of Hindi.

2.2 Conjunct Predicate: A Type of Complex Predicate

Composed of two or more semantic heads, complex predicates are considered as single linguistic entities. It would be a sensitive domain to identify that whether N+V, V+V and Adj+V are complex predicates or just a series of different parts of speech.

A very strong motivation to identify conjunct predicate is its inclusion in lexical semantic resources such as WordNet as a single entry. Being very productive in nature, it is very difficult to come up with an exhaustive list of all conjunct predicates in a language; therefore, some lexical semantic data base is required. Chakrabarti et al. (2007) made use of some constituency tests proposed by Carnie (2012). In their work, noun incorporation in Hindi verbs was confirmed through three constituency tests: addition of accusatives case markers to the noun, movement and addition of modifiers to the noun phrase.

As mentioned in Chakrabarti et al. (2007), Pandharipande (1993) searched for the semantic similarity between V_1 and V_2 in case of Marathi compound predicates. She is quoted to have applied different syntactic phenomenon such as *passivization, participialization, agreement* and *causativization* and concluded that first three are only applied on first Verb whereas the later applied to both verbs in V_{1+} V_2 construction.

Kachru (1993) distinguished compound verbs from V+V sequence and using the clausal derivations drew different categories of compound verbs in case of South Asian language. Semantic compartmentalization of compound verbs of Kalasha based on „prepared" and „unprepared mind" is said to have categories related to the „knowledge" and belief state of the speaker (Bashir, 1993). Several studies including Chakarbarti, Bhattacharyya and Sarma (2007) have contributed towards the distinction of complex predicates from a series of verbs.

3.2 Light Verbs: Semantic and Syntactic Contribution

Final verbs in N+V, Adj +V and V_1+ V_2 constructions are called as light verbs and usually the syntactic head of the construction (Ahmed & Butt, 2011; Kiani, 2013). Jespersen (1949) was said to be the first person who coined the term „complex predicate" which was primarily applied to English V+NP collocation such as *have a nap, take a step, give a birth* etc. As reported in Butt (1995), Hook (1974) provided a list of twenty-four (24) light verbs in Hindi and Urdu. Butt (1995) listed thirteen (13) light verbs for Urdu, and Bukhari (2009) highlighted seventeen (17) light verbs for Gojri. Akhtar (2000) introduced eight (8) light verbs in Punjabi which carry aspectual information in V_1V_2 construction. Furthermore, he claimed that information regarding volitionality can always be carried by Complex predicate in Urdu and Punjabi which made his work eclectic from that of Butt (1995 and 1997). Light verbs are not

semantically void. This phenomenon can easily be understood by realizing the difference between |take a bath‟ and "give a bath‟. These verbs are neither semantically full nor empty, but semantically bleached in some way. Sometimes, it is difficult to recognize them as light verbs. In Compound Predicates (V_1+ V_2), first verb is referred as *polar verb* whereas the second verb is called as *vector verb* (Ahmed, 2010; Ahmed & Butt, 2011; Butt, 2010; Kiani, 2013; Mushtaq, 2015; Schmidt, 1999). Schmidt (1999) is also of the same view that semantic contribution of polar verb is comparatively more than the vector verb which seems to be semantically bleached. Bowern (2010) claimed that light verbs are semantically incomplete or defective and can be categorized based on their syntactic context i.e., proverb. Vector verb is said to lose its original semantics when combined with polar verb in a compound predicate. She tried to present a clear criterion to distinguish polar verb (V_1) from vector verb (V_2). V_2 takes up morphological features in compound predicate; on the other hand, it inflects with tense, aspect and agreement morphology when function as auxiliary. Butt (1995) categorized common light verbs from aspectual and permissive complex predicate as listed below in Table 1.

Table 1 Common Light Verbs (source: Butt, 1995)

Based on (di)transitive (Ergative Subject)	**Based on Intransitives (Nominative Subject)**
le „take‟	*aa* „come‟
de „give‟	*jaa* „go‟
daal „put‟	*par*.„fall‟
maar „hit‟	*mar* „die‟
nikaal „pry out‟	*nikal* „emerge‟
	cuk „finish‟
	baith „sit‟
	uth „rise

Another effort to distinguish between auxiliaries and light verbs has been made by Butt (2010) who believed that light verbs form a distinct syntactic class which differentiates them from auxiliaries and main verbs. Furthermore, light verbs are dependent on another predicative part which is not the case with main verbs as they may cover the complete meaning (Butt, 2010). Light verbs can also be identified and annotated based on manual identification (Hwang et al.,2010).

This research work focuses on the following research questions:

1. What are the possible instances of Urdu conjunct predicates (N+V)?
2. What possible pattern of case marking and volitionality can be predicted on the basis of collocation of Urdu noun and light verb?

3. Research Methodology

As demonstrated in Levin (1993), lexical items with similar meaning tend to exhibit similar syntactic behavior. This theory provides linguistically motivated entries for lexicon verbs or nouns which entail the information of meaning and syntactic expression. As cited in Levin (1993), Bloomfield (1933) referred that lexicon is really an appendix of grammar which is a list of irregularities. According to Bloomfield (1933), lexicon bears minimum information regarding the idiosyncratic behavior of the lexical item; however, Levin (1993) added that the knowledge

possessed by a language speaker regarding a lexical item implies that there is more to lexical knowledge than mere characteristic word-specific features. Surface syntactic forms of arguments and case markers are determined by light verbs in conjunct predicates, whereas nominal assigns semantic roles to the conjunct predicate. Ergative case maker is the sign of agentive role of the subject in perfective aspect of tense i.e., past, present or perfect. There are examples of noun+ light verb instances in Urdu language where ergative case marker "*ne*" does not assign agentive role. In sentence, "*main nay bardashat kia hay*" it assigns experiencer role which can be argued further based on semantic value of related nouns in N+V instances.

According to Levin (1993)'s proposition, behavior of verb such as expression and interpretation of its arguments corresponds to the meaning it carries, can be used to interpret the syntactic behavior of conjunct predicate (N+V) in order to come up with the semantic classes of Urdu nouns. This research ventures to systemize and delimit the aspects of Urdu conjunct predicate.

3.1 Choice of Urdu Corpora

In order to study the semantic classes of nouns in conjunct predicates (N+V), many instances of N+V collocations are required. A huge Urdu corpus was required for this purpose which is fully annotated and most preferably freely available for online use. Urdu Digest Corpus has been developed by Center for Language Engineering (CLE), UET, Lahore (Ehsan & Butt, 2020; Urooj et al., 2012). UrduWaC is a non-tagged 53 million Urdu corpus which consists of Urdu online sources. It was developed by Kilgarriff , Reddy, Pomikálek, and Avinesh (2010) and it is available at Sketch Engine (https://www.sketchengine.eu/urwac-urdu-corpus/). Universal Dependency Urdu Corpus „UD_URDU_UDTB" is a research product of Hindi/Urdu Tree Bank project which is placed at githib.com (Bhat et al., 2017). The rationale for using Universal Dependency (UD) Urdu corpus is many folds. For instance, a broader representation of Urdu dialect used in the world will be incorporated. Lexical tools which will come as a result of this research will be adequately sufficient to train the natural language programs. Apart from its large size and free availability I chose UD Urdu Corpus because of its multilayered annotated nature. It is annotated for morphological features such as gender, number and forms. Subject, object, adjective modifier, auxiliary, etc. have also been annotated. For every word in the sentences; its lemma form has been described with its universal and regional part of speech. It followed the framework of universal dependency for consistent annotation of grammar which makes it unique for its widely practiced technique in the field of natural language processing. The genre of the corpus is news which consists of content from diverse disciplines such as politics, sciences, business, technology, health, society, religion, sports and film industry.

The ultimate motivating factor for using this resource in that "Universal Dependencies Structure" is the tool which has become a first choice of natural language processing experts to develop useful tools in computational linguistics in the recent decade. The tool has been used to extract N+V collocations to inquire the semantic compatibility of nouns with the light verbs in conjunct predicate to anticipate that the resultant research product would contribute to the under resourced computational tools for Urdu.

راؤنڈ	راؤنڈ	NOUN	NNC	Case=Nom|Gender=Masc|Number=Sing|Person=3	2	compound		
ٹیبل	ٹیبل	NOUN	NN	Case=Acc|Gender=Masc|Number=Sing|Person=3	4	nmod	_	
پر	پر	ADP	PSP	AdpType=Post	2	case	_	ChunkId=NP|ChunkType=child
مباحث	مباحث	NOUN	NN	Case=Acc|Gender=Masc|Number=Sing|Person=3	30	obl	_	
سے	سے	ADP	PSP	AdpType=Post	4	case	_	ChunkId=NP2|ChunkType=child
قبل	قبل	ADP	NST	AdpType=Post|Case=Nom|Gender=Masc|Number=Sing|Person=3	4	case		
ہم	ہم	PRON	PRP	Case=Nom|Number=Plur|Person=1|PronType=Prs	30	nsubj	_	
کامن	کامن	PROPN	NNPC	Case=Nom|Gender=Masc|Number=Sing|Person=3	10	compound		
ویلتھ	ویلتھ	PROPN	NNPC	Case=Nom|Gender=Masc|Number=Sing|Person=3	10	compound		
گیمس	گیم	PROPN	NNP	Case=Acc|Gender=Masc|Number=Plur|Person=3	13	nmod	_	
کی	کا	ADP	PSP	AdpType=Post|Case=Nom|Gender=Fem|Number=Sing	10	case	_	
آرگنائزنگ	آرگنائزنگ	ADJ	JJ	Case=Acc	13	amod	_	ChunkId=NP5|ChunkType=child
کمیٹی	کمیٹی	NOUN	NN	Case=Acc|Gender=Fem|Number=Sing|Person=3	16	nmod	_	
کے	کا	ADP	PSP	AdpType=Post|Case=Acc|Gender=Masc|Number=Sing	13	case	_	
برطرف شدہ	برطرف شدہ	ADJ	JJ	Case=Nom	16	amod		ChunkId=NP6|Chu

Figure 1. Urdu_UD_UTB Corpus

4. Result and Discussion

The N+V collocation is extracted by closely monitoring the sentences in UD_URDU_UTB. Multi layered annotation provided an authentic revision to the intuition for picking it up as a light verb. It also provided lemmas for different morphological forms of light verbs which further helped in their clear identification. The most frequent collocated light verbs are identified as *kar, hona, rah, rakh, aa, dia, ja, laga, lee/leya, which* later will incorporate instances of all possible collocated light verbs with nouns in corpus. As elaborated in Table 2, chosen corpus was mined to check the compatibility of different nouns for all collocated light verbs.

Manual extraction of N+V collocations to the extent for reaching some generalizable pattern is laborious; therefore, Lexico-Syntactic Pattern Extraction (LSPE) method is employed to pull all instances of N+V (Hearst, 1998). In LSPE method, first of all, required searchable patterns are manually identified and then an algorithm is formulated to find the pattern in a corpus. Mining complex predicate from a parallel corpus has already been handled in a number of research studies including Sinha (2009).

Corpus has been surveyed to extract nouns and their semantic and syntactic compatibility with the most frequent light verbs is mined, and simultaneously documented using an MS Excel sheet. Chakrabarti et al. (2007) used three constituency tests i.e., addition of accusative case marker to the noun; constituency tests and addition of modifier to the noun phrase in order to decide the N+V, Adj+V and V+V constructions as complex predicate so that these can be inserted in the Urdu WordNet as single entries.

Noun	**Light Verbs**										
	Do /Kar	Become/ *Hu*	Is/ *He*	Remain/ *rah*	*Put/ Rakh*	Come/ *aa*	Give/ *Dia*	Go/ *Ja*	Hit/ *Laga*	Take/ *Lee*	N +V
Memorize/*Yaad*	1	1	1	1	1	1	0	0	0	0	6
Wait/*Intizaar*	1	0	1	0	0	0	0	0	0	0	2
Advise/*Mashwara*	1	0	1	0	0	0	1	0	0	1	4
Demand/*Mutalba*	1	1	1	0	0	0	0	0	0	0	3
Mention/*Zikar*	1	1	1	0	0	0	0	0	0	1	4
Accuse*/Ilzaam*	0	1	1	0	0	0	1	0	1	1	5
Loss*/Khasara*	0	1	1	0	0	0	0	0	0	1	3
Express/*Izhaar*	1	1	1	0	0	0	0	0	0	1	4
Step*/Qadam*	0	0	1	0	1	0	1	0	0	0	3
Claim*/Daaway*	1	1	1	0	0	0	0	0	0	0	3
Commit */Irtikaab*	1	1	0	0	0	0	0	0	0	0	2
Document/*Raqam*	1	1	1	0	0	0	1	0	1	1	6
Chance/*Mouqa*	0	0	1	0	0	0	1	0	1	0	3
End */Anjaam*	1	1	1	0	0	0	1	0	0	0	4
Instruct*/Hadayat*	1	1	1	0	0	0	1	0	0	1	4
Lesson/*Sabaq*	0	1	1	0	0	0	1	0	1	1	5
Light/*Roshni*	1	1	1	0	0	0	1	0	0	1	5
Try/*Koshish*	1	1	1	0	0	0	0	0	0	0	3
Emphasize/*Ahmiat*	1	1	1	0	0	0	1	0	0	1	5
Fight/*Larayai*	1	1	1	0	0	0	0	0	0	1	4
Witness/*Shahadat*	0	1	1	0	0	0	1	0	0	1	4
Use/*Istamaal*	1	1	1	0	0	0	0	0	0	0	3
Poison/*Zehar*	1	1	1	0	0	0	1	0	1	1	6
Disclose*/Inkashaf*	1	1	1	0	0	0	0	0	0	0	3
Position/*Rutba*	0	0	0	0	1	0	1	0	0	1	3
Life/*Zindagi*	0	0	0	0	0	0	1	0	1	1	3
Shame/*Sharam*	1	1	1	0	0	0	0	0	0	0	3
Time*/Waqat*	0	1	1	1	1	1	1	1	1	1	9
Face/*Samana*	1	1	1	0	0	0	0	0	0	0	3
Operation*/Karwai*	1	1	1	0	0	0	0	0	0	0	3
Eradication*/Insidad*	1	1	0	0	0	0	0	0	0	0	2
Decide*/Faisala /Baatcheet*	1	1	1	0	0	1	1	0	0	1	6
Communicate	1	1	1	0	0	0	0	0	0	0	3
Control/*Ikhtiyaar*	1	1	1	1	1	1	0	0	0	1	7
Answer*/Jawab*	0	1	1	0	0	1	1	0	0	1	5
Loss*/Nuqsaan*	1	1	1	0	0	0	1	0	0	1	5
Include/*Shamil*	1	1	1	0	1	0	0	0	0	0	4
Thank/*Tashakar*	1	1	0	0	0	0	0	0	0	0	4
Collide/*Takkar*	1	1	0	0	0	0	1	1	1	1	6
Wound/*Zakham*	1	1	1	1	0	1	1	0	1	1	8
Begin/*Shru*	1	1	1	0	0	0	0	0	0	0	3
Crowd/*Hajoom*	1	1	1	0	0	0	0	0	1	0	4

Table 2. Semantic and Syntactic Compatibility of Nouns with the Light Verbs

As asserted by Ahmed and Butt (2011), light verbs carry information regarding agreement, case marking on subjects and thematic roles i.e., agentive and experiencer. An interesting pattern emerged from the analysis of data that noun in conjunct predicate does not agree with the light verb when preceding argument is marked with nominative case marker, but the noun in N+V construction may agree with the light verb when accusative, instrumental or genitive case marker is used with the inserted argument.

6. Larkay ko sabaq yad hua
Boy-M.Sg=acc lesson M.Sg.Nom. memorized/learnt became-perf.M.Sg.
"Boy learnt the lesson."

7. Larkay nay nazam yad ki
Boy-M.Sg=erg poem F.Sg.Nom. memorized do-perf.M.Sg.
"Boy memorized the lesson."

An extra argument along with the appropriate case marker was inserted to check the gender agreement between noun and light verb within the conjunct predicate (N+V) construction. Bhat (2008) also examined the agreement between noun and light verb with the use of different case makers. Dative, accusative, genitive case markers mostly block the gender agreement with the adjoining argument thus emphasis lies on noun/argument in N+V construction. Light verb hay seemed neutral in all the instances therefore may not be presented as a decisive feature to conclude the final judgment. Testing the pattern of gender agreement between the constituents of conjunct predicate is said to be helpful in deciding that how many of the total recorded N+V instances are conjunct predicates.

Light verbs in Urdu conjunct predicate have their impact on the semantic roles of the subjects as well. Transitive light verbs *kar, rakh* , *dia* and lee allow agentive role of the subject. A total of 1024 extracted instances of Urdu Conjunct predicate have been provided with paper as an Appendix.

5. Conclusion

This analysis helps in drawing patterns of semantic compatibility of noun classes with different light verbs. Based on this, some generalized semantic relevant criteria may be evolved for the identification of conjunct predicates.As mined from the universal dependency Urdu corpus, light verb *'kar'* is the most frequent light verb which collocates in its different morphological forms such as *'ki', 'kia', 'karwaya', 'karain'* with mostly noun communication e.g., *Yad, Wazahat, waada,* and *wazih. etc. Kar* draws ergative case marker „*Ne'* which reflects volitionality of the action on part of the subject.Second most frequent light verb found from the corpus is „*Ho'* which has different inflectional morphological forms e.g., „*Hua'*, „*Hona'*, „*Hoga'*, „*Hona', 'Hoya'* etc. It licenses accusative case marker *'Ko'* which lacks volitionality of the verb.*Rah* with different morphological forms such as *Rahay, Rahta, Raha* allows accusative case marker and collocates with noun state e.g., *Wazih, Mehroom, Mehdood* etc.Patterns of noun classes with collocated light verbs ensure the development of a tool for the identification of conjunct predicates. Theoretical framework used for data collection i.e., Universal Dependencies makes it workable for natural language processing tool development for Urdu which is not fully resourced to date.We focused on all naturally occurring light verbs

in form of conjunct predicate. In a future study, this inventory can be extended by investigating the possible instances of N+V in other Urdu corpora as well I.e., EMILLE /CIIL Urdu (Baker et al., 2002), and Charles University Urdu Monolingual Corpus, (Jawaid et al., 2014). This research yielded the possible instances of Urdu conjunct predicate (N+V) in which different syntactic conditions are examined to reveal the compatibility pattern. Transitive Urdu light verbs which require more than one argument such as *kar*, *rakh*, *dia*, and *lee* license ergative case markers on subject. Furthermore, these light verbs allow accusative case marker '*ko'* on objects and dative case maker '*ko'* on indirect objects. These overall syntactic conditions are found which favor the occurrence of N+V as Urdu conjunct predicate.Nominative case markers which are also called as null case markers are found on the direct argument before the N+V instance which causes light verb to agree with it and hence the data analysis supported the notion that there is no agreement between noun and light verb in N+V instances – this advocates its single syntactic constituency as a conjunct predicate.Semantic correlation can also be claimed based on theoretical analysis of compatibility of transitive light verb with agentive subjects. This Finalizes their status as conjunct predicate along with the accumulated information decoded through the linguistic analysis of light verbs i.e., case marking of subjects, gender agreement and agentive vs. experiencer nature of subjects. As an extension of this work, the next step may be to determine the semantic class of nouns in conjunct predicate.

References

Ahmed, T. (2010). The interaction of light verbs and verb classes of Urdu. *Proceedings of Verb 2010:The Identification and Representation of Verb Features, Pisa, 200,* 10-15.

Ahmed, T., & Butt, M. (2011). Discovering semantic classes for Urdu NV complex predicates. *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011),UK, 9,* 305-309.

Aikhenvald, A.Y. (2006) *Classifiers and noun classes: semantics.* In: Brown, Keith, (ed.) Encyclopedia of languages and linguistics. Elsevier, Oxford, UK, pp. 463-471.

Aissen, J. L., & Perlmutter, D. M. (1983). Clause reduction in Spanish. *Studies in Relational Grammar. 1*(*2*), 360-375.

Akhtar, R. N. (2000). *Aspectual complex predicates in Punjabi.* (Unpublished Doctoral Dissertation). University of Essex, Colchester.

Alsina, A. (1993). *Predicate composition: A theory of syntactic function alternations.* (volumes I and II) Available from ProQuest Dissertations & Theses Global. Retrieved from https://www.proquest.com/dissertations-theses/predicate-composition-theory-syntactic-function/docview/304093479/se-2?accountid=135034

Baker, P., Hardie, A., McEnery, T., Cunningham, H., & Gaizauskas, R. J. (2002, May). EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation. *Proceeding of Language Resource and Evaluation Conference (LERC2002), Las Palmas, Canary Islands,* 819–827.

Bashir, E. (1993). Causal chains and compound verbs. In MK Verma (Ed.*) Complex predicates in south Asian languages*. (pp.1-30). New Delhi: Manohar.

Bhat, R. A., Bhatt, R., Farudi, A., Klassen, P., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., Vaidya, A, Vishnu, S. R., & Xia, F. (2017). The Hindi/Urdu treebank project. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 659–697). Netherlands: Springer. doi:10.1007/978-94-024-0881-2_24

Bloomfield, A. L. (1933). The effect of restriction of protein intake on the serum protein concentration of the rat. *The Journal of experimental medicine, 57*(5), 705.

Bowern, C. (2008). The diachrony of complex predicates. *Diachronica, 25*(2), 161-185.

Bukhari, N. H. (2009). Light verbs in Gojri. *Language in India, 9*(8), 449-463.

Butt, M. (1993). Hindi-Urdu infinitives as NPs. *South Asian Language Review: Special Issue on Studies in Hindi-Urdu, 3*(1), 51-72.

Butt, M. (1993a). Conscious choice and some light verbs in Urdu. In MK Verma (Ed.) *Complex predicates in south Asian languages*. (pp.31-45). New Delhi: Manohar Publishers and Distributors.

Butt, M. (1993b). Hindi-Urdu infinitives as NPs. *South Asian Language Review: Special Issue on Studies in Hindi-Urdu, 3*(1), 51–72.

Butt, M. (1995). *The structure of complex predicates in Urdu*. Center for the Study of Language (CSLI): Stanford University, USA.

Butt, M. (2010). The light verb jungle: Still hacking away. In Mengistu Amberber, Brett Baker & Mark Harvey (Eds.) *Complex Predicates: Cross-Linguistic Perspectives on Event Structure.* (pp.48-78). New South Wales: Cambridge University Press.

Butt, M., & Geuder, W. (2013). On the (semi) lexical status of light verbs. In Norbert Corver & Henk van Riemsdijk (Eds.), *Semilexical categories: On the content of function words and the function of content words*, (pp.323–370). Germany: Walter de Gruyter.

Butt, M., & Ramchand, G. (2001). Complex aspectual structure in Hindi/Urdu. Nomi Erteschik-Shir and Tova Rapoport (Eds.), *The syntax of aspect: deriving thematic and aspectual interpretation* (pp.1-30).Oxford Scholarship Online. doi: 10.1093/acprof/:oso/9780199280445.003.0006

Carnie, A. (2012). *Syntax: A generative introduction* (Vol. 18). John Wiley & Sons.

Cattell, N. R. (1984). *Syntax and semantics: Composite predicates in English*. Academic Press.

Chakrabarti, D., Sarma, V., & Bhattacharyya, P. (2007). Complex predicates in Indian language WordNets. *Lexical Resources and Evaluation Journal*, *40*(3/4). 331-355.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (Eds.). 2020. *Ethnologue: Languages of the World.* 23rd edition. Texas: SIL International. Retrieved from http://www.ethnologue.com.

Ehsan, T., & Butt, M. (2020). Dependency Parsing for Urdu: Resources, Conversions and Learning. *Proceedings of the12th Language Resources and Evaluation Conference, France,* 5202-5207. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.640

Hearst, M. A. (1998). Automated discovery of WordNet relations. In Christiane Fellbaum (Ed.). *WordNet: An Electronic Lexical Database*. (pp.131-151). Cambridge: MIT Press doi**:**https://doi.org/10.7551/mitpress/7287.003.0011

Hook, P. E. (1974). *The compound verbs in Hindi.* Center for South and Southeast Asian Studies: The University of Michigan.

Hook, P. E. (1978). The Hindi compound verb: What it is and what it does. In Kripa Shanker Singh (Ed.). *Readings in Hindi-Urdu linguistics*. (pp.129-154). New Delhi: National Publishing House.

Hwang, J. D., Bhatia, A., Bonial, C., Mansouri, A., Vaidya, A., Xue, N., & Palmer, M. (2010, July). Propbank annotation of multilingual light verb constructions. In *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 82-90).

Jespersen, Otto. 1949. A Modern English Grammar on Historical Principles, Part VI, Morphology. London: George Allen and Unwin Ltd.

Kachru, Y. (1993). Verb serialization in syntax, typology and historical change. In MK. Verma (Ed.) *Complex predicates in south Asian languages.* (pp. 115–134). New Delhi: Manohar Publishers.

Kiani, Z. H. (2013). *The syntax of complex predicates in Urdu*. Lap Lambert. Academic Publishing.

Kilgarriff, A., Reddy, S., Pomikalek, J & PVS, A. (2010). A corpus factory for many languages. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, Malta. 605-619.

Levin, B. (1993*). English verb classes and alternations: A preliminary investigation.* University of Chicago Press.

Mohanan, T. (1994). *Argument structure in Hindi*. Center for the Study of Language (CSLI). Stanford University.

Mohanan, T. (1997). Multidimensionality of representation: NV complex predicates in Hindi. In Alex Alsina, Joan Bresnan & Peter Sell (Eds.). *Complex predicates*. (pp. 431–471). Stanford University: CSLI.

Mushtaq, Z. (2015). *Grammatical analysis of Urdu auxiliary verbs* [Doctoral dissertation]. University of Management and Technology Lahore. Pakistan.

Pandharipande, R. (1993). Serial verb construction in Marathi. In B. D. Joseph & A.M. Zwicky (Eds.). *When verbs collide.* (pp. 178-199). Columbus: Ohio State University.

Poornima, S., & Koenig, J. P. (2009). Hindi aspectual complex predicates. In *Proceedings of the international conference on head-driven phrase structure grammar* (pp. 276-296).

Schmidt, R. L. (1999). *Urdu, an essential grammar*. Routledge.

Sinha, R. M. K. (2009). Mining complex predicates in Hindi using a parallel Hindi-English corpus. *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009), Singapore,* 40–46.

Urooj, S., Hussain, S., Adeeba, F., Jabeen, F., & Parveen, R. (2012). CLE Urdu digest corpus. *L*anguage *& Technology*, 47-54.

Warsi, J. (2004). Conditions under which English is taught in Pakistan: An applied linguistic perspective. *Sarid Journal*, *1*(1), 1-9.

Appendix A

Mined Collocation of N+V from Universal Dependency Urdu Corpus I.

Noun/Adj	LightVerb	Subj	Obj/Obl
یقینی	بنایا		شراکت
یقینی	ہے	کہ	نمائندگی
یقین	دلایا	انہوں	کیا
یرغمال	بنا		سپاہیوں
یاد	رہے	جو	شئے
یاد	رہے		سنائی
یاد	رہے		رکھا
یاد	رہے	جو	کیا
یاد	رہے	وہ	بلا
یاد	رہے	نینس	کی
یاد	رہے	سرینا	دی
یاد	ہوگا	کھمبے	ہو

Noun/Adj	LightVerb	Subj	Obj/Obl
ہدایت	دی		اجلاس
ہدایات	دی	کورٹ	احکام
ہجرت	کریں_گے	تم	باتیں
ہاتھ	جھاڑ	عملہ	ہونے
نوش	فرمایا	حضور	
نوٹ	لیا	پارٹی	
نوٹ	لیا	میں	
نوازا	ہے	تم	کر
نمائندگی	کر	عباس	
نمائندگی	کرتے	اڈوانی	
نگرانی	کریں_گے	رشیدی	
نقل	کر	عوام	