Computational Description of Urdu Collocations

Dr. Mahmood Ahmed Awan¹
Dr. Abrar Hussain Qureshi², PAKISTAN

Abstract: With the rise of the Firthian linguistics, the importance of bringing about a distinct shift from learning and teaching lexical items towards a consideration of lexis in larger units of language organization began to be felt. As meaning in text is cumulative and each lexical item is affected what follows it or precedes it. So the study of lexis does not mean the study of words in isolation. So, collocations play an important role in second language learning and teaching, especially at advance level. There is a slim body of literature on the study of Urdu collocations. The gap is deplorable especially in view of the fact that the use of computer has brought about in the study of language. The EMILLE corpus is a case in point. Employing the EMILLE, the present article aims to explore Urdu collocation vis-à-vis their structure and function. The study shows that collocations play a very significant role in the Urdu language. The finding has pedagogical implications. The teachers and material designers of Urdu must focus on the way collocations create meanings. The article concludes with the hope this article will at a spur to other researchers of Urdu to investigate the manifold aspects of the Urdu collocations.

Key Words: Collocation, description, Urdu

Introduction

There are two main traditions in the study of collocations. The first tradition started with Firth (1951) while the second belonged to Sinclair (1966). The main difference between these two traditions was that Firthian study of collocation was not based on corpus while Sinclair’s was based on corpus. In other words, Sinclair developed the Firthian tradition of collocation with its statistical aspects based on corpus. (Carter, R. & M. McCarthy 1988) observe:

“The central concepts of research done by J. R. Firth (1951) were originally "colligation and collocation". Unfortunately, colligation failed to take off and collocation was recognized and accepted only as an insight. Firth, who was considered by many to be the “Father of collocation”, believed in the separation of lexis and semantics because he thought collocation was the
central part of a word’s meaning. From his research, he was responsible for bringing the term collocation into prominence during the 60’s and the 70’s”.

The tendency of two words to occur together is called ‘collocation’ (Firth 1957) or a pair of words is considered a collocation if one of the words significantly prefers a particular lexical realization of the concept, the other represents. The term “collocation” was made more explicit by Halliday (1961; 1966) that encompasses the statistical distribution of lexical items in context. He observes “Collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at n removes (a distance of n lexical items) from an item x, the items a, b, c…” (Halliday, 1961:276). Benson (1989) states that “collocations should be defined not just as a ‘recurrent word combinations’ but as ‘arbitrary recurrent word combination’. “Arbitrary” as opposed to “regular” means that collocations are not predictable and cannot be translated word by word. According to Sinclair, (1991; 121) collocation is the concept of word co-occurrence, where certain words appear predictably next to or within a certain number of words from each other.

Collocational relations present the greatest problems of replicable identification and that a crucial difficulty remains one of interpreting how loosely or tightly collocating items are strung together (Carter, 1998). Explaining collocation, Lin (1998) observes that even though "baggage and luggage" are synonyms, only "baggage" can be modified by emotional, historical, or psychological considerations. This lack of valid substitution for a synonym is a characteristic of collocation in general (cf. Manning & Schutze 1999) Many have defined these patterns of co-occurrence in the grammatical context that they result primarily from syntactic dependencies or they can be lexical in that, although syntactic relationship is involved. Hoey (2005) defines collocation that it is a psychological association between words (rather than lemmas) up to four words apart and is evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution. This definition is intended to pick up on the fact that collocation is a psycholinguistic phenomenon, the evidence for which can be found statistically in computer corpora.

Nevertheless the definition of the exact nature of a collocation varies from one researcher to the next. It is variously defined as a habitual word combination (Lin 1998) or a recurrent word combination. Of course, if words do regularly collocate in this way, we shall expect to find some semantic relationship among them; but this may be quite complex and indirect. Collocation is a purely lexical relationship; that is, it is an association between one word and another, irrespective of what they mean.

In contrast with collocation, the idea of colligation is that just as a lexical item may co-occur with another lexical item, so also it may occur in or with a particular grammatical function (Hoey, 2005, 43). Fontenelle (1997) defined grammatical collocation that it involved one
element from an open class and an element from a closed class, typically, but not necessarily, a preposition. According to Carter (1998, P.60) collocation is the co-occurrence on nouns, verbs, or adjectives with prepositions or certain features of grammar, for example, a comparative form of an adjective with the word "than", or the verb "deal" with the preposition "with".

Studies of collocation in English, Urdu and many other languages have tended to be within two distinct traditions: one oriented towards specifically grammatical and one towards specifically lexically patterning. The former has tended to result in studies that have been of distinct value to language learners and have to an extent emerged from the demands of particular pedagogic projects. While the tradition of lexical collocation have produced seminal studies which have contributed substantially to our understanding of lexis. According to Carter (1998):

“The processing of collocations involves a number of parameters, the most important of which is the "measure of association", which evaluates whether the co-occurrence is purely by chance or statistically significant. Due to the non-random nature of language, most collocations are classed as significant, and the association scores are simply used to rank the results. Commonly used measures of association include "mutual information, t-score, log-likelihood".

Collocations can be in a syntactic relation (such as object-verb in Urdu), lexical relation (such as antonymy), or they can be in no linguistically defined relation. Some collocations are so strong that it seems strange to list them as separate words, to be combined into the collocation each time we need it.

In order to describe collocations, Sinclair states:

“We may use the term node to refer to an item whose collocations we are studying and, we may define a span as the number of lexical items on each side of a node that we consider revelent to that node. Items in the environment set by the span, will be called collocates” (1966, P.415). Consider the following

**Urdu Collocations**

A recent perspective - corpus based statistical approach - has been adopted here to investigate collocations. As there is no major corpus available of Urdu language like The Bank of English, prepared by University of Birmingham United Kingdom, so, EMILLE, a corpus of Indian language for Bengali, Gujarati, Hindi, Panjabi, Singhalese, Tamil and Urdu, prepared by Lancaster University, United Kingdom, has been used. Though, EMILLE is very small corpus of Indian Languages yet it has been used as a point of departure. “Wordsmith Tools” a software
International Journal of English and Education
ISSN: 2278-4012, Volume:3, Issue:4, October 2014

launched by Oxford University, has been used to analyze concordance and collocational pattern. An example of the collocations of Urdu from the EMILLE corpus of the word سبز is produced below:

In the above concordance, the lexical item سبز sabz is the node word and its collocates have been shown in a span of five on each side of the node.

**Grammatical Collocation**

Grammatical collocation is also called syntactic collocation. According to Biskup, 1992 and Bahns, 1993 grammatical collocations are the type in which a dominant word fits together with a grammatical word, for example, in the following concordance the noun is accompanied by a grammatical word.
The grammatical collocation may also be analysed with the approximate collocates of a grammatical item in the form of a span on each side of the node. Most of the immediate collocates of Urdu words are nouns as in the following concordance:

120 نک آخر میں شروع جوان تو وہ بیوین نہیں علاج لا لگون سے ہے کہ ہیں اس مطلو کا مطلو کے خرچ ہیں کے اس مطلو کے ہیں
121 نفاذ میں علما میں کی آپ میں علاج لا لگون سے ہے کہ ہیں اس مطلو کا مطلو کے خرچ ہیں کے اس مطلو کے ہیں
122 بیماری اس سے متعلقہ کی رہن بیماری کی علاج لا جلد کے نوجوان اس مطلو کا مطلو کے خرچ ہیں کے اس مطلو کے ہیں
123 جب نہیں بیماری والی جائیں معجزہ علاج لا دینا ہے جب سے پتھر ضریح بچ جان سی
124 گا کہ ہیں زمانہ تو وہ اگر ہے محسوس علاج لا لگون سے ہے کہ ہیں اس مطلو کا مطلو کے خرچ ہیں کے اس مطلو کے ہیں
125 نک آخر میں شروع جوان تو وہ بیوین نہیں علاج لا لگون سے ہے کہ ہیں اس مطلو کا مطلو کے خرچ ہیں کے اس مطلو کے ہیں
126 کو بیماری میں ہیں دینے سے ملدی جہاں ہے علاج لا لگون سے ہے کہ ہیں اس مطلو کا مطلو کے خرچ ہیں کے اس مطلو کے ہیں
127 غیر مستقل کی بیوین معجزہ ہے میں علاج لا لگون سے ہے کہ ہیں اس مطلو کا مطلو کے خرچ ہیں کے اس مطلو کے ہیں
128 میں مصرف میں اعمال بیوین چھوٹے علاج لا لگون سے ہے کہ ہیں اس مطلو کا مطلو کے خرچ ہیں کے اس مطلو کے ہیں
129 ان سے را کی ہیں ایک بھی دیدا کی اس مطلو کا مطلو کے خرچ ہیں کے اس مطلو کے ہیں
130 عام کا ہے مطلو کا اس مطلو طالع علاج لا لگون سے ہے کہ ہیں اس مطلو کے خرچ ہیں کے اس مطلو کے ہیں
131 ہے بیماری کا بیماری اور مطلو علاج لا لگون سے ہے کہ ہیں اس مطلو کا مطلو کے خرچ ہیں کے اس مطلو کے ہیں
132 30 نک بھی بھی مطلو کا مطلو کے مطلو کا مطلو کے خرچ ہیں کے اس مطلو کے ہیں
133 ہے بیماری بیماری اور مطلو علاج لا لگون سے ہے کہ ہیں اس مطلو کا مطلو کے خرچ ہیں کے اس مطلو کے ہیں

Lexical collocation may also be called semantic collocation. To attain a clearer understanding of lexical collocations, it is helpful to try to distinguish them from free combinations and idioms. Actually, it is a matter of frequency that constitutes lexical collocations, for example, سز

International Journal of English and Education
ISSN: 2278-4012, Volume:3, Issue:4, October 2014
“sabaz” (green) may collocate with لباس “libaas (clothes), جهندارة “jhanda (flag), maidaan (ground), etc. while free combinations do not occur in language frequently while on the other hand, idioms are relatively frozen expressions whose meanings do not reflect the meanings of their component parts. Lexical collocations do not contain preposition, infinitives, or clauses, but consist of various combinations of nouns, adjectives, verbs, and adverbs (Bahns, 1987). It normally consists of combinations of two basically “equal” words such as noun+verb, noun+noun, adjective+noun etc. in which both words contribute to the meaning.

The above concordance shows that the node word سبز “sabaz” (green) is frequently accompanied by nouns.

Cowie and Howarth (1995) suggest that such lexical phrases can be placed on a 4-level scale of complexity. At Level 1, idioms are multiword lexemes that have frozen collocation. If any variation is inserted into the idiom, it ceases to exist as a unit. In the idiom تارے گننا taaray gin-naa (not to sleep all the night) the meaning would be lost if even one element was changed. Thus idioms are the least complex because they allow no variation. Moving to Level 2, the collocation is still fixed, but the meaning is not idiosyncratic any longer; rather it is more transparent because it is composed from the meanings of all of the component words. This level is somewhat more complex because the meaning has to be composed from several lexemes, rather than coming from a list of words e.g. سر تا پیپر sir ta paer (from head to foot). Level 3 has a slot that can be filled with a limited list of words, most of which are similar in meaning and would be placed at one point of collocation. This choice introduces both variation and complexity e.g. مثال کے قائم کرنا mesaal qaim kerna (to set example) and مثال دینا mesaal deina (to give example) etc. Level 4 has two slots instead of one, adding increased variation and complexity and new words may be placed at the both ends of collocation e.g.:
1. خوشگوار موسم khushgawar----------- mausum
   (pleasant) (weather)

2. ثانی رات Thandi --------------- raat
   (cold) (night)

In the following cluster, the node word پانی paani (water) is placed with a variety of words on either side.

According to Leo Wanner (2002) “to summarise, the idiosyncratic interpretation of the term "collocation" covers all non-free combinations also covered by its distributional interpretation, but, in addition, it covers rare idiosyncratic combinations and rejects all recurrent free combinations. Therefore, it is the interpretation of the term "collocation" as an idiosyncratic combination that underlies the present study. It is to be noted that we make use of the following properties implied by this interpretation.
1. a collocation is a binary combination of lexical items,

2. a collocation possesses a coherent syntactic structure, i.e. the base and the collocate always possess the same grammatical function with respect to each other,

3. a collocation is a lexically restricted word combination, i.e. it cannot be constructed using universal (semantic) selectional restriction rules; rather, the base predetermines the set of lexical items (collocates) it may appear with in a combination based on idiosyncratic, collocation type-specific grounds

In what follows, some collocations from the EMILLE corpus have been analysed in order to see how far they measure up to the above interpretation.

Theoretically, it is possible for any lexical item of Urdu to co-occur or keep company with any other lexical item. However, for any particular lexical item A there are other items which have a high probability of being found near A. For instance, we might accept *نیا* (*naea*) (new) to have a high probability of co-occurrence with *گھر* (*gher*) (house), *کتاب* (*kitab*) (book), *شہر* (*sheher*) (city), etc. but a low probability of co-occurrence with *بیٹری* (*biji*) (electricity), etc. The fact that the patterns are probabilistic raises some methodological problems. For example a fundamental difficulty in measuring collocational relations is that of deciding the maximum distance between items that can be said to be collocating. The solution generally adopted has been a relatively arbitrary of restricting the collocating items to a span of a fixed number of words on either side of the specified focal word whose patterning is being investigated. It is also difficult, however, to appropriately demarcate what might be the upper limit of a lexical unit and to decide whether a larger unit could itself be the limit of the collocational relationship. In Sinclair (1966) and Sinclair et al (1970) the aim was to study large quantities of text in order to focus in a statistically significant way on the company kept by particular words and for the strength and weakness of partnership to be expressed in terms of percentile frequencies of co-occurrence. But not enough data could be processed for
either a wide range of lexical collocational probabilities or interesting lexical sets to be described. The corpus used in the present study is too small to solve this thorny issue. A corpus larger in fairly size may do so.

In Urdu and many other languages of the world some words are relatively freer than other in forming pairs, e.g. اچھا (good) can have freer combination with many nouns اچھا لڑکا (good boy), etc. Some words occur only in company of some specific class of words. For instance, the adjective با سی (not fresh) makes a collocation like با سی سبزی baasi sabzi (stale vegetable) but does not make a collocation with some uncooked edibles like گہوں gehu (ghee) as با سی گہوں baasi gehu. In a semantic field the relationship between the two words is paradigmatic forming a system of vertical substitution. In a collocation the relationship between the two words is syntagmatic.

A full description of the collocational behaviour of even a small group of lexical items requires the analysis of a great number of texts from different registers. The practical problems, such studies used to encounter, can now be partially overcome with the development of computational linguistics and the use of computers in linguistic analysis, and particularly in frequency counts. The analysis of larger samples will possibly give more valuable information about how language works at the level of lexis. In the first place collocational descriptions will indicate the different syntagmatic relationships; a certain lexical item enters into in different contexts of situation. The occurrences of a certain item on the syntagmatic axis are part of the meaning.

**TYPES OF COLLOCATIONS IN URDU**

Following are the main types of collocations in Urdu

1. Unrestricted collocation: This describes the capacity of particular lexical items to be open to partnership with a wide range of item. Most core words fall into such category. For example پانی paani (water) collocates relatively unrestricedly with a wide range of animate and inanimate, concrete and abstract entities: e.g. پانی لگانا panni lagaana (to water), پانی پینا paani peena (to drink water), پانی پنا panni daalnaa (to throw water), پانی پنیلی paani peelana (to make some one drink), پانی ملنا paani meelana (to mix water), etc. There are,
however, ‘transitional areas’ (Cruse, 1986 p.41) between free combinations/collocations and collocations/idioms. The fact may be proved with the help of corpus that the node word *paani* (water) accompany with a big slot of words.

**ii. Semi-restricted collocation**

This category embraces lexical patterns in which the number of items, which can be substituted in different syntactic slots, is more determined. For example, the collocations of *aasmaani* (heavenly) are comparatively restricted e.g. *aasmaani katub* (heavenly books), *aasmaani rhematae* (heavenly blessings), *aasmaani aafteen* (heavenly sufferings)
iii. Restricted Collocation: These restricted are based neither on semantic range nor or meaning but confined to one or two words only e.g. *chichilati dhoop* (scorching sun), *chiknie chupri baate* (flattering words), *gora rung* (white complexion). In other words, these collocations can appear only in the company of words of a specific class and not with others.

iv. Inherent collocation: Some words inherently require a human participant and some inhuman, and some either of them, similarly, some words are used in the context of animate participants and some of inanimate, and some of both. For instance, *rona: ‘to weep; represents a universal human activity*
PATTERNS OF LEXICAL COLLOCATIONS IN URDU

1. Noun+Noun Collocation

N Concordance

13  کر مار نے اس کا آخر جلن؟ تک کہ دھونا رونا کر کہ۔ مگر وہ رے زندگی کی کرتوں سمجھتے۔
14  ہیں خوشی بھی ہے۔ نہیں ہے صرف دھونا رونا کر کہ۔ چاہئے اور دوسرے ہیں۔ کہ انسان کی قسم میں
15  پروفیسر۔ بیان جلی۔ گھر اب چھوڑنا دھونا رونا۔ نہیں کہ چاہئے پر پناہ بھی دھونا رونا کر کہ
16  بات۔ نموہار سادات ایک وہ رونا وقت اسی تھی رہی رہے رونا اور قطر رونا۔ گھی جسن وقت وہ اپنی تقیید کا
17  نگو والی رہی میں ملبے کے اس رہتا۔ رونا رونا کر نہ رہن حرام بوگرتی۔ وہ اکثر اپنی بدنخستگی کا
18  حقوق گنتے جوہیں س نے تمیاری بیس۔ رونا رونا تھا گنتا بیس۔ تمہارے پہلے وہ وقت اپنی بہت ماری کا
19  مجبور ہے۔ اعلی میں آب پہلیا۔ سکیا لاس وہ نہیں جیسے سکا کر۔ چہ کرم نے سکا اور
20  دو لمحے۔ اسی آب بے طور بوگرتی۔ سکیا لاس رونا گھی (52) 17 پہلی رونا۔ اور
21  نبی، رونا سے دوڑ روز پہنچانے۔ سکیا لاس رونا کر ایک تھا۔ بہ تو میں چہ نہیں۔ سکا۔
22  اپنے وہ جب بابر حالات پر۔ دو نہیں کر شروع رونا سے اور اب بہ تو میں چہ #دو #کے سیٹی بی

2. Noun+Verb Collocation

13  چتر نہا کے بیس کے گھر رونے جے ایک میں گھر بریض کو دیکھنے گیا۔ وہ
14  کبیا کا اشکا موجودہ۔ کو فریب کے گھر پہنچ واقع 30 اکتوبر 1883 کو پس شہ آئے۔ ان کے
15  پہلے بے لوٹنے اپنے پر گھر پہنچ کے سمجھے بہی پہنچ رہا۔ دیکھنے گیا۔ چاہے
16  چتر نہا کے بیس کے گھر رونے جے ایک میں گھر بریض کو دیکھنے گیا۔ وہ
17  کبیا کا اشکا موجودہ۔ کو فریب کے گھر پہنچ واقع 30 اکتوبر 1883 کو پس شہ آئے۔ ان کے
18  پہلے بے لوٹنے اپنے پر گھر پہنچ کے سمجھے بہی پہنچ رہا۔ دیکھنے گیا۔ چاہے
19  چتر نہا کے بیس کے گھر رونے جے ایک میں گھر بریض کو دیکھنے گیا۔ وہ
20  کبیا کا اشکا موجودہ۔ کو فریب کے گھر پہنچ واقع 30 اکتوبر 1883 کو پس شہ آئے۔ ان کے
21  پہلے بے لوٹنے اپنے پر گھر پہنچ کے سمجھے بہی پہنچ رہا۔ دیکھنے گیا۔ چاہے
22  اپنے وہ جب بابر حالات پر۔ دو نہیں کر شروع رونا سے اور اب بہ تو میں چہ #دو #کے سیٹی بی
According to Allerton (1984) verb-noun constructions in which the verb carries little semantic content and is used for the sake of its structural properties only (1984) as in

3. Adjective + Noun collocations.

According to Allerton (1984) verb-noun constructions in which the verb carries little semantic content and is used for the sake of its structural properties only (1984) as in
1. Adverb + adjective collocations: Some adverbs modify particular types of adjectives.

Concluding Remarks

Although it is widely acknowledged that collocations are both indispensable and problematic for language learners and that they therefore should play an important role in teaching and learning second language, especially at an advanced level. Study of the Urdu language at the level of Collocation is a new phenomenon. But owing to its significance, It has great importance. The present study tried to investigate collocations in Urdu language in detail yet, the study is far from being comprehensive in view of the fact that the size of the corpus used in this study was small. A larger corpus would have made this study more comprehensive in which more dimensions might have been investigated in the end. Investigation of collocations of the Urdu language is a vast and interesting field. The authors of this article hope that this research will motivate linguists to explore more dimensions of Urdu collocations in more detail.

References


Cowie, A.P. (1978) “The place of illustrative material and collocations in the design of a learner’s dictionary. Firth (1951) "

Fontenelle, T. (1997) *Turning A Bilingual Dictionary Into Lexical-Semantics Database*. Published in cooperation with the dictionary Society of North America (DSNA) and the European Association for lexicography (EURALEX)


First Author: Dr. Mahmood Ahmed Awan. Associate Professor Govt.College Burewala. Pakistan

Second Author: Dr. Abrar Hussain Qureshi, Assistant Professor of English Govt. College Chichawatni Pakistan