Volume 4, Issue 11, November– 2026 **ISSN (E):** 2938-379X

THEORETICAL AND METHODOLOGICAL ISSUES IN THE CORPUS-BASED STUDY OF UZBEK COLLOCATIONS

Radjabova Gulnoza Giyosiddinovna
PhD in Pedagogical Sciences, Associate Professor
Uzbekistan State University of World Languages
E-mail: rad.gulnoza@gmail.com
Tel: +998 971007901

Abstract
This article examines the theoretical and methodological problems of identifying collocations in Uzbek by means of corpus methods. It focuses on the status of compound verbs, on lemmatisation and on word order.

Keywords: Collocation, corpus linguistics, compound verb, phraseological unit, semantic prosody, lemmatisation, mutual information, national corpus of Uzbek.

Introduction

ANNOTATSIYA
Maqolada o‘zbek tilidagi kollokatsiyalarni korpus asosida aniqlashning nazariy va metodologik muammolari, xususan, qo‘shma fe’llar maqomi, lemmatizatsiya va so‘z tartibi masalalari yoritiladi.

Kalit so‘zlar: kollokatsiya, korpus lingvistikasi, qo‘shma fe’l, frazeologik birlik, semantik prosodiya, lemmatizatsiya, o‘zaro axborot ko‘rsatkichi, o‘zbek tili milliy korpusi.

АННОТАЦИЯ
В статье рассматриваются теоретические и методологические проблемы корпусного изучения узбекских коллокаций: статус составных глаголов, лемматизация и порядок слов.

Ключевые слова: коллокация, корпусная лингвистика, составной глагол, фразеологическая единица, семантическая просодия, лемматизация, мера взаимной информации, национальный корпус узбекского языка.

INTRODUCTION

A speaker of Uzbek will say yuksak mukofot (‘high award’) and baland ovoz (‘loud voice’, literally ‘high voice’), but would hardly say baland mukofot or yuksak ovoz. Both adjectives express height, and all four phrases are grammatically well formed. Even so, native speakers hear two of them as natural and the other two as odd. This feeling comes neither from a

Web of Teachers: Inderscience Research
webofjournals.com/index.php/1
Licensed under a Creative Commons Attribution 4.0 International License.

grammatical rule nor from the dictionary meaning of the words. It rests on a habit built up over many years of exposure to which words tend to appear together. Modern linguistics calls this phenomenon collocation.

The study of fixed expressions has a rich tradition in Uzbek linguistics. Sh. Rahmatullayev's explanatory dictionary of phraseology [2] and A.E. Mamatov's work on lexical and phraseological norms [1] are important milestones in it. Phraseological research, however, has understandably concentrated on idioms with a figurative meaning. The combinations that sit between the free phrase and the phraseological unit, which are transparent in meaning but not freely assembled, have not yet been studied systematically on corpus evidence, or so it seems to us. The reason is easy to see. Such combinations can be listed by intuition, but to measure how stable they are one needs a large collection of texts, that is, a corpus.

The topic is timely because work on a national corpus of Uzbek has gathered pace in recent years, and an international conference has even been devoted to it [4]. Once a corpus becomes widely available, a new question faces the linguist: what can actually be learned from it, and how must it be built for the findings to be reliable? The aim of this article is to clarify the theoretical status of collocations in Uzbek and to set out the methodological conditions for studying them on corpus data.

LITERATURE REVIEW AND METHODOLOGY

The idea that the meaning of a word is tied to its habitual "neighbours" is much older than corpus linguistics. According to J.R. Firth's frequently quoted remark, a word can be known by the company it keeps [6, 11]. More than thirty years later J. Sinclair, working with the first large electronic corpora, concluded that text is built on two principles. Under the open-choice principle, any slot in a sentence may be filled by any word the grammar allows. The idiom principle holds that a speaker has at their disposal a large store of semi-preconstructed phrases, each of which is realised as a single choice [13]. In Sinclair's view, it is the second principle that dominates ordinary text.

Sinclair later put forward the model of the "extended unit of meaning", in which a lexical item is described through four kinds of relation: collocation (the words it occurs with), colligation (the grammatical patterns it combines with), semantic preference (the semantic set of words it tends to select) and semantic prosody (the evaluative colouring it takes on) [14]. M. Stubbs developed the model further on English material [15]. What matters for us is that a collocation is not just a pair of words but the visible part of a larger pattern.

In corpus linguistics, statistical measures are used to decide whether words co-occur by chance. The most widespread of them is mutual information (MI), which K. Church and P. Hanks introduced into lexicography [5]. Pairs with an MI value of 3 or above are usually treated as collocations [7]. MI, however, overrates rare but tightly bound pairs, while the t-score does the opposite and puts frequent combinations first. For this reason researchers recommend using the two measures together [7; 9].

Uzbek phraseology has traditionally relied on V.V. Vinogradov's classification. Of its three types (phraseological fusions, phraseological unities and phraseological combinations), the one closest to collocation is the phraseological combination, because in it the meaning of one

Web of Teachers: Inderscience Research
webofjournals.com/index.php/1
Licensed under a Creative Commons Attribution 4.0 International License.

component is realised only within a restricted range of words. This tradition, though, has tended to regard such combinations as a separate, peripheral layer of the vocabulary. Corpus studies have shown that they are in fact very common in running text [13].

The study is theoretical and analytical in nature and relies on descriptive, comparative and componential analysis. The material consists of common word combinations in contemporary standard Uzbek. As far as the author is aware, there is at present no large, balanced Uzbek corpus that offers open access to collocational statistics and could be used by an individual researcher. For that reason the Uzbek examples below are presented not as quantitative evidence but as hypotheses to be tested in a corpus in the future.

DISCUSSION AND RESULTS

Free combinations, collocations and phraseological units

Word combinations can be arranged along a line. At one end stand free combinations such as yangi kitob (‘new book’) and katta hovli (‘big yard’). In these, any component can be replaced by any word of suitable meaning: yangi kitob, eski kitob, qiziqarli kitob (‘new, old, interesting book’). At the other end are idioms such as tarvuzi qo‘ltig‘idan tushmoq (literally ‘for one’s watermelon to fall from under one’s arm’, i.e. to lose heart). Their meaning does not follow from the words they contain, and a language learner has no choice but to memorise them.

Collocations lie between these two poles. In combinations such as yuksak mukofot (‘high award’), qattiq tanqid (‘harsh criticism’), chuqur hurmat (‘deep respect’) and qizg‘in bahs (‘heated debate’), the base word (mukofot, tanqid, hurmat, bahs) keeps its literal meaning. The second component, however, is not freely chosen, and its meaning, most often one of intensity or high degree, is realised only within that particular combination. Qizg‘in with bahs denotes not heat but vigour, and yet one cannot say qizg‘in hurmat to express the same intensity.

The main mark of a collocation is therefore a peculiar asymmetry: its meaning is clear, but its form cannot be predicted. A foreigner learning Uzbek understands qattiq tanqid at once but would find it hard to produce the phrase unaided, since several adjectives, such as kuchli (‘strong’), keskin (‘sharp’), jiddiy (‘serious’) and qattiq (‘hard’), all seem logically suitable. Which of them are really used, and which is the most frequent, can be shown only by long experience or by a corpus. This is exactly what makes collocations one of the hardest parts of language learning and translation.

Compound verbs as collocations

Uzbek grammar treats compound verbs, which are formed from a noun and an auxiliary verb (qilmoq, etmoq, bo‘lmoq and others), as a separate category: harakat qilmoq (‘to act’), xizmat qilmoq (‘to serve’), qabul qilmoq (‘to accept’), tashakkur etmoq (‘to thank’). From the grammatical point of view, each of them is a single lexeme. From the point of view of collocation the question is a different one. Why does ta’sir (‘influence’) combine with qilmoq, etmoq and ko‘rsatmoq, yordam (‘help’) with bermoq and ko‘rsatmoq, and e’tibor (‘attention’) with bermoq and qaratmoq? The choice of the verb is lexical, not grammatical. In this respect it is no different from the choice of verb in the English pay attention or exert influence.

Web of Teachers: Inderscience Research
webofjournals.com/index.php/1
Licensed under a Creative Commons Attribution 4.0 International License.

Many compound verbs, then, are lexical units and collocations at the same time, and the boundary between the two is gradual, not sharp. A corpus makes it possible to measure these degrees, and at least three criteria can be used for the purpose. The first is the variability of the verb component, that is, whether the noun combines with a single verb or with several. The second is whether other words can come between the components. One can ask how often, alongside ta'sir ko'rsatdi ('had an influence'), forms such as ta'sir ham ko'rsatdi ('also had an influence') or ta'sirini sezilarli darajada ko'rsatdi ('showed its effect to a considerable degree') are found. The third is whether the noun takes the accusative suffix. In Uzbek an unmarked object forms a tighter unit with the verb, whereas an object with -ni points to a definite, known referent and is syntactically freer. In this sense the ratio of qaror qabul qildi ('made a decision') to qarorni qabul qildi ('accepted the decision') can serve as a kind of gauge of how far a combination has become integrated.

The table below groups combinations that a speaker's intuition treats as synonymous but whose range of use remains to be checked in a corpus.

Table 1 Synonymous combinations and directions for corpus testing

Combinations	**By intuition**	**Question to test in the corpus**
ta'sir qilmoq / ta'sir etmoq / ta'sir ko'rsatmoq ('to influence')	near-synonyms, hardly any difference	distribution across styles and genres; which adjectives (large, negative, positive) they occur with
yordam bermoq / yordam ko'rsatmoq ('to help')	the same meaning	degree of combination with modifiers such as medical, legal, material
e'tibor bermoq / e'tibor qaratmoq ('to pay attention')	freely interchangeable	tendency towards official-administrative or journalistic texts
qaror qabul qilmoq / qaror chiqarmoq ('to make a decision')	partly different	type of subject: a person or an institution (court, council)
baland / yuksak + noun ('high')	both express "height"	ratio of combination with concrete and abstract nouns

Each row of the table advances a particular assumption based on intuition. The fact that tibbiy yordam ko'rsatish ('to provide medical assistance') has become established as a term, for instance, supports the assumption that yordam ko'rsatmoq leans towards institutional, professional help, while yordam bermoq leans towards everyday situations. If the corpus confirms this, then for dictionaries and textbooks the two combinations are not plain synonyms but units with different ranges of use. If it does not, the intuition was mistaken. Either result is useful, since both turn a guess into evidence. This is also why the table gives no figures: such numbers can come only from a corpus, and putting in rough estimates would repeat the very error this article criticises.

Web of Teachers: Inderscience Research
webofjournals.com/index.php/1
Licensed under a Creative Commons Attribution 4.0 International License.

Web of Teachers: Inderscience Research
webofjournals.com/index.php/1

Semantic prosody and hypotheses for Uzbek

Semantic prosody refers to the evaluative colouring that a word "catches" from its habitual collocates. In English, for example, the verb set in has been found to occur mainly with subjects that denote unpleasant states of affairs, although its dictionary meaning contains nothing negative [13]. Features of this kind are almost invisible to intuition. A native speaker uses the word correctly but can seldom explain what colouring it carries [15].

Uzbek offers no shortage of examples that point to the same phenomenon. In a speaker's intuition, sodir etmoq ('to commit') attaches almost exclusively to nouns such as crime, offence and murder. Its relative sodir bo'lmoq ('to occur') is used both with unpleasant events, as in yo'l-transport hodisasi sodir bo'ldi ('a road accident occurred'), and in neutral sentences such as jiddiy o'zgarishlar sodir bo'ldi ('serious changes took place'). Two combinations from a single root, in other words, may differ in the strength of their prosody. Yo'l qo'ymoq ('to allow', 'to let happen'), when it takes a noun in the dative case, seems to select words such as error, shortcoming, violation of the law and negligence, which is why a sentence like yutuqlarga yo'l qo'ydi ('allowed achievements') sounds strange. Avj olmoq ('to intensify', 'to flare up') occurs with nouns such as disease, epidemic and conflict in medical and social texts, but in texts describing work processes it appears in neutral combinations like qurilish ishlari avj oldi ('construction work gathered pace'). This suggests that its prosody may depend on genre.

All of these observations remain, for now, hypotheses based on intuition. They are precisely the kind of question that has to be tested in a corpus, since research on English has shown that semantic prosody is among the features least reflected in speakers' intuition [14; 15]. Explanatory dictionaries give the meaning of a word but do not record its evaluative colouring in any systematic way. For translators and learners of Uzbek, however, this is exactly the information that is needed.

Methodological requirements for an Uzbek corpus

To be able to answer the questions above, a corpus must meet at least five conditions.

The first is lemmatisation. The verb ko'rsatmoq ('to show') occurs in running text in dozens of forms: ko'rsatdi, ko'rsatgan, ko'rsatayotganini, ko'rsata olmadi, ko'rsatilmoqda and so on. A search by word form scatters the data, and MI calculated on rare forms comes out artificially high. Collocation statistics for Uzbek must therefore be computed at the level of lemmas. Voice forms are a separate matter. Passive forms such as qaror qabul qilindi ('a decision was made') and yordam ko'rsatildi ('help was provided') are very common in official texts. If the lemmatiser does not link qilinmoq to qilmoq, a large share of collocations will be left out of the count.

The second condition is the correct handling of multi-word units and of the context window. A corpus should allow compound verbs to be searched both as whole units and as separate components. In addition, in Uzbek the object comes before the verb and the modifier before the word it modifies. In the sentence Bu qaror mintaqa iqtisodiyotiga sezilarli ta'sir ko'rsatdi ('This decision had a considerable effect on the regional economy'), sezilarli stands to the left of ta'sir, and ta'sir to the left of ko'rsatdi. The right-hand context span used in studies of English

Licensed under a Creative Commons Attribution 4.0 International License.

verb–object combinations does not give the expected results for Uzbek, where the decisive context lies on the left.

The third is spelling normalisation. Uzbek texts exist in parallel in Latin and Cyrillic script (ta'sir and таъсир). Even within the Latin script, the letters o‘ and g‘ and the apostrophe (tutuq belgisi) are typed with several different characters (o‘, o‘, o’, o` and so on). Because the character codes differ, a computer may count the single word ko‘rsatdi as three or four different words. Frequency figures from a corpus whose spelling has not been normalised cannot be trusted.

The fourth is genre and style metadata. Most of the hypotheses in the table, as well as the assumption about avj olmoq, concern style and genre. A corpus that does not separate literary, scientific, official, journalistic and conversational texts cannot answer such questions.

The fifth is parallel corpora. Comparative research and translation practice need parallel corpora of Uzbek with English and Russian. Automatic text alignment is a particular difficulty here. G.G. Radjabova's study of English–Russian parallel corpora found that the largest share of alignment errors, 24 per cent, stemmed from differences in word order [11]. Uzbek word order differs from English more consistently than from Russian, and the language is agglutinative, so the problem can be expected to be even sharper for the Uzbek–English pair.

Once these conditions are met, a corpus-based collocational dictionary of Uzbek becomes possible. In such a dictionary each headword would be given with its most typical collocates, ranked by frequency and statistical strength, together with examples and genre labels. It would complement Sh. Rahmatullayev's phraseological dictionary [2] and the five-volume explanatory dictionary [3], which quote combinations as examples but say nothing about their frequency.

Such a corpus could be used in teaching as well. T. Johns's data-driven learning has learners work directly with corpus evidence [8]. G.G. Radjabova's studies on English material have described what corpus technologies can offer in teaching academic writing [10] and how they help B2-level students improve their writing competence [12]. When an open Uzbek corpus appears, similar activities could be introduced in native-language lessons. Students could, for example, analyse concordance lines and work out for themselves how yordam bermoq differs from yordam ko‘rsatmoq.

CONCLUSION

The analysis leads to the following conclusions. First, collocation in Uzbek is a distinct phenomenon lying between the free word combination and the phraseological unit, and its main marks are transparency of meaning and unpredictability of form. Second, many compound verbs are collocations as well. Their degree of cohesion can be measured by the variability of the verb component, by whether other words may come between the components, and by whether the noun takes the accusative suffix. Third, there are examples suggesting that semantic prosody exists in Uzbek too, but for now they remain hypotheses based on intuition. Fourth, for a national corpus to be able to answer these questions, it needs lemmatisation (including voice forms), handling of multi-word units and left-hand context, spelling normalisation and genre metadata.

Web of Teachers: Inderscience Research
webofjournals.com/index.php/1
Licensed under a Creative Commons Attribution 4.0 International License.

A corpus does not make the linguist's intuition redundant. Reading and interpreting concordance lines remains the researcher's task. What a corpus changes is the status of intuition, which moves from being a judge to being a source of hypotheses. The questions raised in this article are offered as a research agenda to be tested once the national corpus of Uzbek is opened for general use. Some of them will be confirmed and some refuted. Either way, Uzbek linguistics stands to gain.



Web of Teachers: Inderscience Research
webofjournals.com/index.php/1

REFERENCES

1. Mamatov A.E. Hozirgi o‘zbek adabiy tilida leksik va frazeologik norma muammolari (in Uzbek). – Toshkent, 1991. – 212 b.
2. Rahmatullayev Sh. O‘zbek tilining izohli frazeologik lug‘ati (in Uzbek). – Toshkent: O‘qituvchi, 1978.
1. O‘zbek tilining izohli lug‘ati. Besh jildlik (in Uzbek). – Toshkent: O‘zbekiston milliy ensiklopediyasi, 2006–2008.
2. O‘zbek tilining milliy korpusi: muammo va vazifalar: Xalqaro ilmiy-amaliy konferensiya materiallari (2022-yil 31-may) (in Uzbek). – Toshkent: ToshDO‘TAU, 2022. – 356 b.
3. Church K.W., Hanks P. Word association norms, mutual information, and lexicography // Computational Linguistics. – 1990. – Vol. 16, No. 1. – P. 22–29.
4. Firth J.R. A synopsis of linguistic theory, 1930–1955 // Studies in Linguistic Analysis. – Oxford: Blackwell, 1957. – P. 1–32.
5. Hunston S. Corpora in Applied Linguistics. – Cambridge: Cambridge University Press, 2002.
6. Johns T. Should you be persuaded: Two samples of data-driven learning materials // English Language Research Journal. – 1991. – Vol. 4. – P. 1–16.
7. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. – Cambridge: Cambridge University Press, 2012.
8. Radjabova G.G. Corpus technologies in teaching academic writing // O‘zbekistonda xorijiy tillar. – 2023. – № 1 (48). – B. 92–103.
9. Radjabova G.G. Linguistic aspects of automatic text alignment in parallel corpora // Western European Journal of Linguistics and Education. – 2025. – Vol. 3, Issue 5. – P. 121–125.
10. Radjabova G.G. Using Corpus of Contemporary American English [COCA] to enhance writing competence of B2 level students // O‘zbekistonda xorijiy tillar. – 2025. – Vol. 11, № 6 (65). – B. 110–124.
11. Sinclair J. Corpus, Concordance, Collocation. – Oxford: Oxford University Press, 1991.
12. Sinclair J. The search for units of meaning // Textus. – 1996. – Vol. 9, No. 1. – P. 75–106.
13. Stubbs M. Words and Phrases: Corpus Studies of Lexical Semantics. – Oxford: Blackwell, 2001.



Licensed under a Creative Commons Attribution 4.0 International License.