STUDYING UZBEK COLLOCATIONS THROUGH CORPUS DATA: THEORETICAL FOUNDATIONS AND METHODOLOGICAL CHALLENGES

Radjabova Gulnoza Giyosiddinovna

PhD in Pedagogical Sciences, Associate Professor

Uzbekistan State University of World Languages

E-mail: rad.gulnoza@gmail.com

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.*

ABSTRACT

This article looks at the theoretical and methodological problems of identifying collocations in Uzbek with corpus methods, in particular the status of compound verbs, lemmatisation and word order.

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

INTRODUCTION

A speaker of Uzbek will say *yuksak mukofot* (‘high award’) and *baland ovoz* (‘loud voice’), but would hardly say *baland mukofot* or *yuksak ovoz*. Both adjectives express height and all four phrases are grammatical, yet native speakers hear two as natural and two as odd. The feeling comes from a habit built up over years of hearing which words tend to go together, not from a grammatical rule or the dictionary meaning. Modern linguistics calls this phenomenon collocation.

Uzbek linguistics has a rich tradition of studying fixed expressions, represented by Sh. Rahmatullayev’s phraseological dictionary [2] and A.E. Mamatov’s work on lexical and phraseological norms [1]. Understandably, though, phraseological research has focused on idioms with a figurative meaning. Combinations that stand between the free phrase and the idiom, transparent in meaning but not freely assembled, have not yet been studied systematically on corpus evidence. Intuition can list such combinations, but only a corpus can measure how stable they are. As work on a national corpus of Uzbek gathers pace [4], the question is how it must be built for the findings to be reliable. This article aims to clarify the theoretical status of Uzbek collocations and to set out the conditions for studying them on corpus data.

LITERATURE REVIEW AND METHODOLOGY

The idea that a word's meaning is tied to its habitual "neighbours" is much older than corpus linguistics: J.R. Firth held that a word can be known by the company it keeps [6, 11]. Working with the first large corpora, J. Sinclair concluded that text rests on two principles: open choice, under which any slot can be filled by any word the grammar allows, and the idiom principle, under which speakers draw on a large store of semi-preconstructed phrases, each realised as a single choice [13]. In ordinary text the second dominates. He later proposed the "extended unit of meaning", which describes a lexical item through *collocation* (the words it occurs with), *colligation* (the grammatical patterns it combines with), semantic preference and semantic prosody (the evaluative colouring it takes on) [14; 15]. A collocation is therefore not just a pair of words but the visible part of a larger pattern.

Whether words co-occur by chance is judged with statistical measures. The most widespread is mutual information (MI) [5], and pairs with an MI of 3 or above are usually treated as collocations [7]. MI overrates rare but tightly bound pairs and the t-score favours frequent ones, so both are best used together [7; 9]. In V.V. Vinogradov's classification, long used in Uzbek phraseology, the type closest to collocation is the phraseological combination, yet it has been treated as a peripheral layer of the vocabulary although corpus studies show it is very common in running text [13].

The study is theoretical and analytical, using descriptive, comparative and componential analysis of common combinations in contemporary standard Uzbek. As far as the author is aware, no large, balanced Uzbek corpus with open collocational statistics is available to an individual researcher, so the examples below are hypotheses to be tested in a corpus, not quantitative evidence.

DISCUSSION AND RESULTS

Free combinations, collocations and idioms

Word combinations can be arranged along a line. At one end stand free combinations such as *yangi kitob* (‘new book’), where any component can be replaced by a word of suitable meaning (*eski kitob, qiziqarli kitob*). 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), whose meaning does not follow from their words and which a learner simply has to memorise.

Collocations lie between these poles. In *yuksak mukofot* (‘high award’), *qattiq tanqid* (‘harsh criticism’), *chuqur hurmat* (‘deep respect’) and *qizg‘in bahs* (‘heated debate’) the base word keeps its literal meaning, but the second component is not freely chosen, and its meaning, usually one of intensity, is realised only in that combination. *Qizg‘in* with *bahs* denotes vigour, not heat, yet one cannot say *qizg‘in hurmat* to express the same intensity. The mark of a collocation is thus a peculiar asymmetry: its meaning is clear, but its form cannot be predicted. A foreigner understands *qattiq tanqid* at once but would struggle to produce it, since *kuchli*, *keskin* and *jiddiy* seem equally suitable; hence collocations are hard for learners and translators.

Compound verbs as collocations

Uzbek grammar treats compound verbs, formed from a noun and an auxiliary (*qilmoq, etmoq, bo‘lmoq*), such as *qabul qilmoq* (‘to accept’) or *tashakkur etmoq* (‘to thank’), as single lexemes. From the collocational viewpoint one may ask 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 verb is lexical, not grammatical, as in the English *pay attention*.

Many compound verbs are thus lexical units and collocations at once, and a corpus can measure the degree by three criteria. The first is the variability of the verb component. The second is whether other words can come between the components, for example how often *ta'sir ko'rsatdi* ('had an influence') appears as *ta'sir ham ko'rsatdi* ('also had an influence'). The third is whether the noun takes the accusative suffix: an unmarked object forms a tighter unit with the verb than an object with *-ni*, which is definite and syntactically freer. The ratio of *qaror qabul qildi* ('made a decision') to *qarorni qabul qildi* ('accepted the decision') can thus gauge how integrated a combination is.

Table 1 groups combinations that intuition treats as synonymous but whose range of use remains to be checked.

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	distribution across styles and genres; typical adjectives
yordam bermoq / yordam ko'rsatmoq ('to help')	the same meaning	combination with modifiers such as medical, legal, material

Each row advances an assumption based on intuition. Because *tibbiy yordam ko'rsatish* ('to provide medical assistance') is an established term, one may suppose that *yordam ko'rsatmoq* leans towards institutional help and *yordam bermoq* towards everyday situations. A corpus will either confirm this

or show the intuition wrong, and both results turn a guess into evidence. The table gives no figures for the same reason: they can come only from a corpus.

Semantic prosody

Semantic prosody is the evaluative colouring a word "catches" from its habitual collocates: the English verb *set in*, for instance, occurs mainly with unpleasant states of affairs although its dictionary meaning is neutral [13]. Speakers rarely notice such colouring [15].

Uzbek offers similar examples. *Sodir etmoq* ('to commit') attaches almost exclusively to nouns like crime or murder, while its relative *sodir bo'lmoq* ('to occur') appears both with unpleasant events (*yo'l-transport hodisasi sodir bo'ldi*, 'a road accident occurred') and in neutral sentences (*jiddiy o'zgarishlar sodir bo'ldi*, 'serious changes took place'). For now these remain hypotheses based on intuition, precisely the kind that must be tested in a corpus, since semantic prosody is among the features least reflected in speakers' intuition [14; 15]. Dictionaries do not record it systematically, though translators and learners need it.

Requirements for an Uzbek corpus

To answer such questions, a corpus must meet at least five conditions.

The first is lemmatisation. The verb *ko'rsatmoq* ('to show') occurs in dozens of forms (*ko'rsatdi*, *ko'rsatgan*, *ko'rsatilmoqda* and so on). A search by word form scatters the data, and MI computed on rare forms is artificially high, so collocation statistics must be calculated on lemmas. Passives such as *qaror qabul qilindi* are common in official texts, so the lemmatiser must link *qilinmoq* to *qilmoq*.

The second is the handling of multi-word units and of the context window. Compound verbs must be searchable both as wholes and as separate components. In Uzbek the object precedes the verb and the modifier precedes the modified

word: in *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 span used for English verb–object combinations therefore does not work here; the decisive context lies on the left.

The third is spelling normalisation. Uzbek texts exist in both Latin and Cyrillic script (*ta'sir* and *таъсир*), and even within Latin the letters o' and g' and the apostrophe are typed with several different characters. A computer may then count *ko'rsatdi* as several different words.

The fourth is genre and style metadata, since most hypotheses above concern style and genre.

The fifth is parallel corpora with English and Russian. Automatic alignment is difficult: in G.G. Radjabova's study of English–Russian corpora, 24 per cent of alignment errors, the largest share, stemmed from word order [11]. Uzbek, an agglutinative language whose word order differs consistently from English, should be harder still.

Once these conditions are met, a corpus-based collocational dictionary of Uzbek becomes possible, with typical collocates ranked by frequency and strength. It would complement the dictionaries [2; 3], which give no frequencies, and suit data-driven learning [8; 10; 12], where students analyse concordance lines themselves.

CONCLUSION

The analysis leads to four conclusions. First, collocation in Uzbek is a distinct phenomenon between the free combination and the idiom, marked by transparency of meaning and unpredictability of form. Second, many compound verbs are also collocations, and their cohesion can be measured with the three criteria above. Third, semantic prosody in Uzbek remains an intuition-based

hypothesis. Fourth, a national corpus can answer these questions only with lemmatisation, left-hand context, spelling normalisation and genre metadata.

A corpus does not replace the linguist's intuition: interpreting concordance lines remains the researcher's task. It only changes the status of intuition, from judge to source of hypotheses. The questions raised here form a research agenda for when the national corpus of Uzbek opens for general use; some will be confirmed, some refuted, and either way Uzbek linguistics gains.

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.
3. O‘zbek tilining izohli lug‘ati. Besh jildlik (in Uzbek). – Toshkent: O‘zbekiston milliy ensiklopediyasi, 2006–2008.
4. 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.
5. Church K.W., Hanks P. Word association norms, mutual information, and lexicography // Computational Linguistics. – 1990. – Vol. 16, No. 1. – P. 22–29.
6. Firth J.R. A synopsis of linguistic theory, 1930–1955 // Studies in Linguistic Analysis. – Oxford: Blackwell, 1957. – P. 1–32.
7. Hunston S. Corpora in Applied Linguistics. – Cambridge: Cambridge University Press, 2002.
8. Johns T. Should you be persuaded: Two samples of data-driven learning materials // English Language Research Journal. – 1991. – Vol. 4. – P. 1–16.

9. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. – Cambridge: Cambridge University Press, 2012.

10. Radjabova G.G. Corpus technologies in teaching academic writing // O‘zbekistonda xorijiy tillar. – 2023. – № 1 (48). – B. 92–103.

11. 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.

12. 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.

13. Sinclair J. Corpus, Concordance, Collocation. – Oxford: Oxford University Press, 1991.

14. Sinclair J. The search for units of meaning // Textus. – 1996. – Vol. 9, No. 1. – P. 75–106.

15. Stubbs M. Words and Phrases: Corpus Studies of Lexical Semantics. – Oxford: Blackwell, 2001.