DATA-DRIVEN LEARNING OF ENGLISH COLLOCATIONS FOR UZBEK-SPEAKING STUDENTS: USING FREE ONLINE CORPORA IN THE UNIVERSITY CLASSROOM

Rasulov Ibrohimjon Akramzoda
Teacher, Philology,
Turan International University,
Namangan, Uzbekistan
overcome2792@gmail.com
0009-0000-9533-9637
https://doi.org/10.5281/zenodo.24133093

Abstract. Collocation errors such as do a mistake, give an exam or drink medicine persist in the English of advanced Uzbek-speaking students long after most grammar errors have disappeared. Many of them can be traced to direct translation from Uzbek or Russian. Corpus linguistics offers a way of tackling them: data-driven learning (DDL), in which learners examine concordance lines to discover how words are actually used. A meta-analysis has shown DDL to be effective, yet corpus methods are nearly absent from English teaching publications in Uzbekistan. This article reviews research on collocation and DDL, classifies typical miscollocations of Uzbek learners by their L1 source, and proposes a four-step DDL lesson cycle (Notice, Query, Generalise, Use) based on free tools: SkELL, the corpora at english-corpora.org and the Compleat Lexical Tutor. A paper-based version for classrooms without internet access is also described.

Keywords: collocation, data-driven learning, corpus linguistics, concordance, lexical approach, L1 transfer, Uzbek learners

O‘ZBEK TILIDA SO‘ZLASHUVCHI TALABALARGA INGLIZ TILI KOLLOKATSIYALARINI KORPUS ASOSIDA O‘RGATISH: UNIVERSITET AUDITORIYASIDA BEPUL ONLAYN KORPUSLARDAN FOYDALANISH

Annotatsiya. Maqolada o‘zbek tilida so‘zlashuvchi talabalarning ingliz tilidagi so‘z birikmalari (kollokatsiyalar) bilan bog‘liq xatolari va ularni korpusga asoslangan o‘qitish (data-driven learning) orqali bartaraf etish yo‘llari ko‘rib chiqiladi. Xatolar ona tili va rus tilidan bevosita tarjima bilan bog‘liq holda tasniflanadi. Bepul onlayn korpuslar asosida to‘rt bosqichli dars sikli (payqash, so‘rov, umumlashtirish, qo‘llash) va internet bo‘lmagan sharoit uchun qog‘oz variant taklif etiladi.

Kalit so‘zlar: kollokatsiya, korpus lingvistikasi, konkordans, leksik yondashuv, interferensiya

ОБУЧЕНИЕ АНГЛИЙСКИМ КОЛЛОКАЦИЯМ УЗБЕКОЯЗЫЧНЫХ СТУДЕНТОВ НА ОСНОВЕ КОРПУСНЫХ ДАННЫХ: ИСПОЛЬЗОВАНИЕ БЕСПЛАТНЫХ ОНЛАЙН-КОРПУСОВ В ВУЗОВСКОЙ АУДИТОРИИ

Аннотация. В статье рассматриваются ошибки узбекоязычных студентов в английских коллокациях и пути их преодоления с помощью обучения на основе корпусных данных. Предлагается четырёхэтапный цикл занятия с использованием бесплатных корпусных инструментов и его вариант на бумажных конкордансах для аудиторий без доступа к интернету.

Ключевые слова: коллокация, корпусная лингвистика, конкорданс, лексический подход, интерференция

INTRODUCTION

When teachers mark essays by English majors in Uzbekistan, they find fewer grammar errors in the third year than in the first, but collocation errors remain: I did many mistakes, We will give our exam tomorrow, You should drink this medicine, There was a strong rain. Such errors rarely get in the way of understanding, but they immediately mark the writer as a non-expert user and make the text look weaker than it is.

Collocation, the tendency of words to co-occur, has been central to linguistics since Firth [4] proposed that a word is known by the company it keeps. Corpus research [1] showed how extensive this tendency is, and Lewis [5] built a teaching approach on it. Learner corpus studies confirm that collocation is a persistent problem even for advanced learners and that many errors are linked to the first language [2, 3].

One promising response is data-driven learning (DDL), introduced by Johns [6], in which learners consult corpus data directly. Free online corpora and learner-friendly interfaces now make DDL possible without specialist software. Nevertheless, corpus-based teaching is almost absent from English teaching publications and practice in Uzbekistan. This article classifies typical collocation errors of Uzbek-speaking learners by source and proposes a practical DDL procedure built on free tools.

LITERATURE REVIEW

Collocation and learners

Sinclair [1] distinguished the open-choice principle, by which any word that fits the grammar can fill a slot, from the idiom principle, by which speakers draw on semi-preconstructed phrases. Much of natural, fluent text is produced by the idiom principle. Lewis [5] argued that teaching should focus on such chunks rather than on isolated words and grammar rules.

Nesselhauf [2] analysed verb–noun collocations in essays by advanced German learners and found that a substantial share were deviant and that the first language was involved in many of the errors. Laufer and Waldman [3] compared learner and native writing and found that learners used far fewer verb–noun collocations and that errors persisted across proficiency levels. Both studies suggest that collocation does not take care of itself through exposure; it needs explicit attention.

Data-driven learning

Johns [6] described the learner as a “researcher” who examines concordance lines, short extracts showing a word in context, and discovers patterns. Instead of being told that we make mistakes, learners see many lines with make a mistake and none with do a mistake, and draw the conclusion for themselves. Boulton and Cobb [7] meta-analysed 64 empirical DDL studies and found large average effects, both in comparisons with control groups and in pre-test/post-test designs, with benefits for vocabulary and lexico-grammar in particular.

Tools

Full corpus interfaces can overwhelm learners. Kilgarriff, Marcowitz, Smith and Thomas [10] described SkELL (Sketch Engine for Language Learning), a free interface that shows simple example sentences and “word sketches”, lists of a word’s typical collocates organised by grammatical relation. The corpora at english-corpora.org, including the Corpus of Contemporary American English [8], allow registered users a limited number of free queries per day. The Compleat Lexical Tutor [9] offers free concordancers and vocabulary profilers used widely in teaching.

MATERIALS AND METHODS

The study combines error classification with teaching design. Typical collocation errors were compiled from the author's experience of marking student writing and grouped by probable source: Uzbek, Russian, or intralingual (confusion within English). For each error, the Uzbek and Russian expressions were identified to show the transfer route. Free corpus tools were then evaluated for accessibility, interface simplicity and suitability for B1–C1 learners. Finally, a lesson cycle was designed by combining DDL principles [6, 7] with collocation teaching principles [5, 11]. The error list is illustrative and should be confirmed with a learner corpus.

RESULTS

Typical miscollocations and their sources

Table 1 lists typical errors with their likely source.

Table 1. Typical collocation errors of Uzbek-speaking learners (illustrative)

Learner form	Target form	Uzbek	Russian	Probable source
do a mistake	make a mistake	xato qilmoq	сделать ошибку	qilmoq covers both do and make
give an exam	take an exam	imtihon topshirmoq ('hand over')	сдавать экзамен	Uzbek and Russian ('hand in')
drink medicine	take medicine	dori ichmoq	пить лекарство	Uzbek and Russian
strong rain	heavy rain	kuchli yomg'ir	сильный дождь	Uzbek and Russian
bitter tea (meaning strong)	strong tea	achchiq choy	крепкий чай	Uzbek
make a photo	take a photo	rasmga olmoq	сделать фото	Russian
read a lecture	give a lecture	ma'ruza o'qimoq	читать лекцию	Uzbek and Russian ('read')
high-qualified specialist	highly qualified specialist	yuqori malakali	высококвалифицированный	Uzbek and Russian
receive knowledge	gain/acquire knowledge	bilim olmoq	получать знания	Uzbek and Russian ('receive')

Two things stand out. First, many errors have support from both Uzbek and Russian, which makes them especially stable: when two known languages agree, learners have little reason to doubt the pattern. Second, the most frequent problem involves delexical verbs (make, do, take, give, have), which carry little meaning of their own, so learners choose them by translating rather than by sense.

Evaluation of free tools

Table 2. Free corpus tools for classroom DDL

Tool	Strength	Limitation	Best use
SkELL	Simple sentences, word sketches, no registration	Fewer search options	B1–B2 learners checking collocates
english-corpora.org (COCA, BNC)	Large, balanced corpora; frequency data; genre comparison	Registration; daily query limit	Teachers preparing materials; C1 learners
Compleat Lexical Tutor	Concordancer, vocabulary profiler	Older interface	Teachers preparing paper concordances

The Notice–Query–Generalise–Use cycle

The lesson cycle has four steps and can be completed in about 40 minutes of an 80-minute class.

1. Notice (10 minutes). Learners receive five sentences from anonymised student essays, each with one miscollocation. In pairs they guess which words are "not quite English".

2. Query (10 minutes). Pairs search SkELL for the key noun (e.g. mistake) and examine the word sketch for verbs used with it as an object. They record the top five verbs and two example sentences.

3. Generalise (10 minutes). Pairs formulate a rule in their own words, for example: "We make a mistake, a decision, a plan, progress; we do homework, research, business." The class compares rules and the teacher adds the observation that the Uzbek qilmoq covers both.

4. Use (10 minutes). Learners correct the original sentences and write three new sentences about their own experience using the collocations found. Correct collocations are added to a class collocation list.

After several lessons, learners can use the Query step on their own when revising their writing. This is the long-term goal of DDL: learners who check usage for themselves.

A paper-based version

Where internet access in classrooms is unreliable, the teacher prepares concordance printouts in advance using the Compleat Lexical Tutor or english-corpora.org. Each printout contains 15–20 lines centred on the node word. The Notice, Generalise and Use steps remain the same; the Query step becomes a reading task in which pairs underline collocates and count them. Johns's [6] original materials were themselves printed concordances, so this version is not a poor substitute.

DISCUSSION

The classification suggests that the most persistent collocation errors among Uzbek-speaking learners are those supported by both Uzbek and Russian. Traditional correction, in which the teacher crosses out do and writes make, rarely changes such patterns, because the learner's two known languages continue to support the error. DDL changes the source of authority: learners see that, in thousands of English texts, make a mistake is common and do a mistake almost absent. Evidence of this kind persuades more than a teacher's red pen, and the same search can be repeated for any word.

The approach also has a cross-linguistic component. In the Generalise step, comparing the English pattern with Uzbek qilmoq and Russian сделать helps learners understand why the error occurs. This links DDL with contrastive awareness, a combination that suits multilingual Uzbek learners.

The approach has three limitations. First, DDL requires training: learners need two or three guided lessons before they can interpret word sketches independently. Second, corpora reflect American or British usage, and teachers should decide which variety is the course norm. Third, the error list in this article is illustrative. A learner corpus of essays by Uzbek university students, annotated for collocation errors, would allow a quantitative study of frequency and source and would provide authentic material for the Notice step.

CONCLUSION

Collocation errors in the English of Uzbek-speaking students are frequent, persistent and in many cases supported by both Uzbek and Russian. Data-driven learning, now possible with free online tools, gives learners evidence that can override these patterns and a method for checking usage independently. The Notice–Query–Generalise–Use cycle, in an online or paper-based version, offers a practical way to bring corpus methods into university English courses in Uzbekistan. A learner corpus of Uzbek students' writing would be a valuable next step, for research and teaching alike.

REFERENCES

1. Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.
2. Nesselhauf, N. (2005). Collocations in a learner corpus. John Benjamins.
3. Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. Language Learning, 61(2), 647–672.
4. Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In Studies in linguistic analysis (pp. 1–32). Blackwell.
5. Lewis, M. (1993). The lexical approach: The state of ELT and a way forward. Language Teaching Publications.
6. Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. English Language Research Journal, 4, 1–16.
7. Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. Language Learning, 67(2), 348–393.
8. Davies, M. (2008–). The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/
9. Cobb, T. (n.d.). Compleat Lexical Tutor [Online tools]. https://www.lextutor.ca
10. Kilgarriff, A., Marcowitz, F., Smith, S., & Thomas, J. (2015). Corpora and language learning with the Sketch Engine and SKELL. Revue française de linguistique appliquée, 20(1), 61–80.
11. Nation, I. S. P. (2013). Learning vocabulary in another language (2nd ed.). Cambridge University Press.