THE ROLE OF CORPUS EVIDENCE IN PHRASEOLOGICAL UNIT PEDAGOGY

Abdukadirova Nasiba Alimjanovna, GSPI, Associate professor (PhD)
e-mail: abdukadirovanasiba2@gmail.com
Abdurazakova Shahida Rahimovna, GDU, Associate professor (PhD)
e-mail: abdurazakovashahida24@gmail.com
Tuychibayev Ikrom Bahodir ugli, GSPI, Teacher
e-mail: ikromtuychibaev96@gmail.com

Abstract. Phraseological units – such as collocations, idioms, and lexical bundles – are critical for achieving language fluency. Corpus linguistics provides a data-driven basis for identifying and teaching these multi-word expressions, revealing their frequency, usage patterns, and contextual meanings. This paper explores how theories like Sinclair's idiom principle and Hoey's lexical priming inform language pedagogy, with special attention to collocational competence in second-language learners. The implications for instructional design and learner autonomy are also discussed.

Keywords: corpus linguistics, phraseological units, collocations, idioms, lexical priming, language teaching, second language acquisition, learner corpora**,** fluency, formulaic language;

FRAZEOLOGIK BIRLIKLARNI O‘QITISHDA KORPUS MA’LUMOTLARINING O‘RNI

Annotatsiya. Frazeologik birliklar – kollokatsiyalar, idiomalar va leksik bog‘lamlar – til ravonligiga erishishda muhim ahamiyatga ega. Korpus lingvistikasi ushbu ko‘p so‘zli iboralarni aniqlash va o‘rgatishda ma’lumotlarga asoslangan yondashuvni taklif qiladi; u ularning chastotasi, qo‘llanish uslublari va kontekstual ma’nolarini ochib beradi. Ushbu maqolada Sinclairning “idioma prinsipi” va Hoeyning “leksik tayyorlik” (lexical priming) nazariyalarining til o‘qitishdagi ahamiyati o‘rganiladi. Ayniqsa, ikkinchi til o‘rganuvchilarida kollokatsion kompetensiyaga alohida e’tibor qaratiladi. Ta’lim dizayni va o‘quvchi mustaqilligiga ta’siri ham muhokama qilinadi.

Kalit so‘zlar: korpus lingvistikasi, frazeologik birliklar, kollokatsiyalar, idiomalar**,** leksik priming**,** til o‘qitish**,** ikkinchi tilni o‘rganish**,** o‘quvchi korpuslari**,** ravonlik**,** formulaviy til;

РОЛЬ КОРПУСНЫХ ДАННЫХ В ПРЕПОДАВАНИИ ФРАЗЕОЛОГИЧЕСКИХ ЕДИНИЦ

Аннотация. Фразеологические единицы – такие как коллокации, идиомы и лексические блоки – играют ключевую роль в достижении беглости речи. Корпусная лингвистика предоставляет основанную на данных основу для выявления и преподавания этих многословных выражений, раскрывая их частотность, особенности употребления и контекстуальные значения. В данной статье рассматривается, как теории, такие как принцип идиомы Синклера и лексическое праймирование Хоуи, влияют на методику преподавания языка.

Особое внимание уделяется формированию коллокационной компетенции у изучающих второй язык. Также обсуждаются последствия для проектирования учебных программ и развития автономии учащихся.

Ключевые слова: корпусная лингвистика, фразеологические единицы, коллокации, идиомы, лексическое праймирование, преподавание языка, изучение второго языка, учебные корпуса, беглость речи, формульный язык;

Introduction. Phraseological units (PUs) play a fundamental role in language proficiency, offering learners ready-made, semantically unified expressions that streamline communication. However, their non-compositional meanings and usage constraints pose challenges for second-language (L2) learners. Traditional grammar-based instruction, which emphasizes rule-based learning and individual words, often fails to address these complexities. The emergence of corpus linguistics -the analysis of large language data collections – has revolutionized how PUs are identified and understood, stressing their centrality in authentic discourse.

Phraseological units include various multi-word expressions such as collocations, idioms, lexical bundles, and formulaic sequences that function as single semantic units. The emergence of corpus linguistics has revolutionized our understanding of these linguistic phenomena and their importance in language pedagogy . Corpus linguistics, which involves the analysis of large collections of authentic language data, has revealed that language is much more patterned and formulaic than previously believed.

While traditional grammar instruction often focused on individual words and rules, corpus linguistics has highlighted the prevalence and pedagogical value of phraseological units in real communication. For example, learners are more likely to encounter and remember phrases like “take into account” or “at the end of the day” than to construct them from individual words. Unlike isolated vocabulary items, phraseological units carry pragmatic and cultural meanings, often making them difficult to translate directly. In comparison to more analytic approaches, corpus-informed methods offer a more realistic model of how language is used in context. This shift emphasizes the need for language teaching to incorporate more natural and frequent multi-word expressions, moving beyond sentence-level grammar to discourse-level competence.

According to Wray, formulaic language makes up between 30% and 50% of discourse, highlighting the central role of phraseology in language use. This understanding has profound implications for language teaching, as traditional word-by-word approaches fail to address the ubiquitous nature of phraseological units in authentic communication.

This research carries important implications for language teaching, as traditional approaches that focus on individual words and grammatical rules often fail to reflect how language is used authentically. Emphasizing phraseological units in teaching not only enhances learners’ fluency and communicative competence but also bridges the gap between classroom language and real-world usage. Therefore, incorporating these units into instructional practices is essential for more effective and realistic language acquisition.

The theoretical framework underpinning corpus-informed approaches to teaching phraseological units draws from several complementary perspectives:

Sinclair’s idiom principle posits that language users have available to them a large number of semi-preconstructed phrases that constitute single choices in the lexicon. This contrasts with the open-choice principle, which views language production as a series of slots to be filled with individual words.

John Sinclair proposed two fundamental principles to explain how language is structured and processed: *the idiom principle* and *the open-choice principle*. These concepts are central to corpus linguistics and the study of phraseology.

The idiom principle asserts that fluent language users frequently rely on a mental repository of semi-preconstructed phrases – also known as formulaic sequences or multi-word expressions – that are treated as single units of meaning and usage. These include collocations, idioms, lexical bundles, and other recurrent word combinations. According to Sinclair, speakers do not construct every utterance from scratch but instead retrieve these ready-made units, which speeds up processing and enhances communicative efficiency. These phrases are not always idiomatic in the traditional sense (e.g., "kick the bucket") but may be regular and transparent combinations (e.g., "make a decision," "in the middle of").

The idiom principle reflects the chunking nature of language use, where prefabricated segments are employed rather than word-by-word construction. This principle is strongly supported by corpus-based evidence showing that natural discourse is highly patterned and repetitive.

In contrast, *the open-choice principle* conceptualizes language as a slot-and-filler system, in which grammatical structures offer open slots that can be filled by any word class that is syntactically appropriate. This approach assumes a more generative and compositional model of language, emphasizing flexibility and creativity.

While the open-choice principle is necessary to account for novel utterances and syntactic variation, Sinclair argued that it is secondary to the idiom principle in actual language use. In practice, speakers default to the idiom principle, especially in fluent and spontaneous communication, because it reduces cognitive load and enhances fluency.

These principles have significant *implications* for:

- **Language Teaching:** Instruction should focus more on teaching chunks and collocations, not just isolated vocabulary and grammar.
- **Lexicography:** Dictionaries must include common phraseological patterns, not just single-word entries.
- **Natural Language Processing (NLP):** Algorithms for machine translation and language modeling benefit from incorporating phraseological units for greater accuracy.

Hoey's theory of lexical priming suggests that words become "primed" to occur in particular contexts and co-occurrences through repeated exposure. This explains why certain word combinations sound natural while others, though grammatically correct, may sound odd to native speakers.

Collocations are words that frequently co-occur with statistical significance. Nation classifies them along a continuum from free combinations to restricted collocations. Corpus research by Nesselhauf has shown that even advanced learners struggle with collocational competence, particularly with combinations that are semantically transparent but lexically restricted.

Nadja Nesselhauf investigates how second language (L2) learners – especially those at advanced levels – deal with *collocational competence*. Collocations refer to combinations of words that frequently co-occur in natural language, and these combinations are often *semantically transparent* (i.e., their meaning is clear from the individual words) but lexically restricted (i.e., only specific words can be used together). Nesselhauf's research sheds light on the difficulties L2 learners face when acquiring such combinations.

Collocations are often classified into several categories, such as:

Adjective + Noun (e.g., strong coffee, heavy rain)

Verb + Noun (e.g., make a decision, take a risk)
Noun + Noun (e.g., traffic jam, data analysis)
Adverb + Adjective (e.g., deeply concerned, highly motivated)

For native speakers, the use of these combinations feels *natural* and intuitive, as they have been learned through extensive exposure to the language. However, for L2 learners, mastering collocations can be challenging because of the lexical restrictions involved. Even if a learner understands the individual meanings of the words involved, they may still struggle with knowing which words can be combined together naturally. For instance, in English, it is common to say *strong coffee*, but it would sound odd to say *powerful coffee,* even though strong and *powerful* are synonyms in some contexts.

Nesselhauf's corpus-based research has shown that even advanced L2 learners often struggle with collocational competence, particularly when dealing with combinations that are semantically transparent but lexically restricted. This finding is significant because it challenges the assumption that advanced learners have acquired full fluency in collocations by the time they reach higher proficiency levels.

The research suggests that while advanced learners may have a solid grasp of grammar and vocabulary, they still face considerable difficulty in using collocations accurately. This is because collocations are not simply a matter of combining words based on their individual meanings but also involve *linguistic intuition* about which words naturally occur together in a particular language. This intuition is shaped by repeated exposure to language in context and cannot be easily acquired through explicit learning of individual words or grammatical rules.

One of the main issues identified in Nesselhauf's work is the phenomenon of overuse of non-native collocations, where learners mistakenly produce word combinations that may be grammatically correct but do not reflect the authentic usage patterns of native speakers. For example, a learner might say do a mistake instead of the correct make a mistake, or strong tea instead of weak tea, despite understanding the individual meanings of strong and weak.

Additionally, advanced learners may fail to recognize that some collocations have a preference for particular lexical items, even when alternative words are semantically close. For instance, learners might be aware that strong can be used in contexts such as strong argument or strong opinion, but might not know that powerful is a more natural choice in other contexts, like powerful speech or powerful influence.

Nesselhauf's research highlights the usefulness of corpus linguistics in identifying these patterns. By analyzing large, authentic corpora of learner language, Nesselhauf was able to uncover the specific areas where advanced learners struggle with collocations and identify the most frequent errors. This approach allowed her to pinpoint particular lexical pairs that learners often misuse, which can then be targeted in language teaching.

For instance, in a corpus of learner language, a phrase like make a photo might appear frequently, even though native speakers would typically say take a photo. Such insights are valuable for both lexical acquisition and teaching methodologies.

Nesselhauf's findings have significant implications for language pedagogy:

Collocation Instruction: Teaching collocations should go beyond isolated vocabulary lists and grammar rules. Learners need to be exposed to frequent word combinations in context to develop their collocational competence

Corpus-Based Pedagogy: The use of learner corpora can provide teachers with real data on the types of errors learners make and the most problematic areas in their use of collocations. Teachers can then design targeted activities to address these issues.

Authentic Input: Teachers should expose learners to authentic language input, such as spoken corpora or real-world texts, to provide examples of how collocations are used naturally.

Phraseological units (PUs) – which include idioms, collocations, fixed expressions, and phrasal verbs – play a vital role in mastering a language. They convey figurative meaning, social tone, and cultural references. However, learners often struggle with acquiring these units due to their non-literal meanings and usage restrictions. Corpus-informed teaching offers a solution by providing real, context-rich examples from authentic language data. Using corpora helps learners understand how PUs function in everyday discourse, improving both receptive and productive skills.

Phraseological units are multi-word expressions whose meanings cannot always be deduced from their individual components. For instance, the idiom:

"Kick the bucket"

This phrase does not literally mean to kick a bucket. In corpus examples, it appears in sentences such as:

"He finally kicked the bucket last winter after a long illness."

Here, the idiom means to *die.* Corpus data shows it is typically used informally and often appears in narratives or humorous contexts. Teaching this through a corpus makes learners aware of its connotative tone, frequency, and contexts where it's appropriate or inappropriate.

Corpora highlight how certain words naturally occur together, a feature known as collocation. For example:

"Make a decision" vs. "Do a decision"

The corpus shows that "make a decision" occurs far more frequently than "do a decision," which is rare or unnatural. Learners benefit from understanding:

"The board made a decision to cut funding."

This helps reinforce correct usage. Without corpus data, learners may rely on logical assumptions, leading to errors.

By displaying multiple corpus lines with the same idiom, learners can see its grammatical flexibility and typical use. Consider:

"Spill the beans"

Corpus lines may show:

"He accidentally spilled the beans about the surprise party."

"Don't spill the beans before the announcement!"

Learners notice that the verb form changes, but the idiom retains its meaning – to reveal a secret. They also observe that the idiom is often used in informal or spoken contexts.

Corpora allow teachers to show how the usage of PUs varies across genres. For instance:

"In hot water"

"The minister found himself in hot water after the controversial decision."

The phrase means to be in trouble. Corpus data shows it is frequent in journalistic writing about public figures. Thus, learners understand not only meaning but who typically uses it and where.

High-frequency PUs should be prioritized in teaching. Corpus tools provide ranked lists of idioms and collocations based on usage frequency. Consider:

"Take into account"

"We must take into account all possible outcomes."

Corpus data shows this is much more frequent than alternatives like "consider into," which do not exist in natural usage. Presenting this data helps learners focus on real, productive expressions.

Some PUs allow for variation, while others are fixed. Corpus use clarifies this:

"Hit the nail on the head"
"You really hit the nail on the head with that analysis."

This idiom is very fixed in structure. Corpus evidence shows that changes like *"hit the hammer on the nail"* are virtually nonexistent and would confuse listeners. Understanding this helps learners avoid incorrect rewording.

Some idioms carry cultural meaning that corpus examples help decode. For instance:

"Bite the bullet"
"I finally bit the bullet and went to the dentist."

The idiom originates from historical wartime practices but is now used metaphorically to mean endure something painful. Corpus usage shows it is often employed in first-person narratives and with tasks people avoid.

Allowing learners to explore corpus examples themselves (e.g., via tools like COCA or BNC) builds analytical skills. Students can:

- Identify which verb forms co-occur with an idiom.
- Note collocational partners (e.g., "strong tea" vs. "powerful tea").
- Distinguish formal vs. informal expressions.

This leads to discovery-based learning, where students become active participants in the process of language acquisition.

Conclusion:

Corpus linguistics, through its empirical analysis of authentic language data, provides a reliable foundation for identifying frequent and contextually appropriate phraseological patterns. Sinclair's idiom principle and Hoey's lexical priming theory offer valuable insights into how language is cognitively processed and stored, affirming the psychological and functional reality of formulaic sequences in fluent discourse. In contrast to the open-choice principle, which presumes syntactic freedom and compositional creativity, these corpus-based theories reveal that real-world language use heavily relies on prefabricated chunks that facilitate processing efficiency and ease of communication.

The evidence presented, particularly from Nesselhauf's studies, highlights that even advanced second language (L2) learners encounter persistent difficulties with collocational competence, especially in managing semantically transparent but lexically restricted combinations. This indicates that lexical knowledge alone is insufficient without a deep familiarity with authentic usage patterns. Collocational errors, such as do a mistake instead of make a mistake, reflect learners' overgeneralization based on grammatical logic rather than exposure to conventionalized usage.

These findings carry significant pedagogical implications. First, they call for a shift in instructional focus from individual word teaching to the teaching of frequent and contextually relevant word combinations. Second, they support the integration of corpora both general and learner-specific into the language classroom to provide learners and instructors with data-driven insights into authentic usage. Such integration enables the identification of common learner errors and supports the development of targeted teaching materials that reflect actual language use.

In sum, corpus-informed pedagogy offers a powerful framework for enhancing the teaching and acquisition of phraseological units. By aligning instruction with the realities of language as it is used, rather than as it is traditionally taught, this approach fosters greater fluency, accuracy, and appropriateness in learner production. As language education continues to evolve, the incorporation of corpus-based methodologies represents not merely an innovation, but a necessary progression toward more effective and realistic models of language learning.

Corpus-informed approaches to teaching phraseological units bridge the gap between theory and practice by showing how language is used in real communication. Rather than teaching idioms or collocations as isolated phrases, corpora reveal their frequency, register, grammatical behavior, and discourse functions. This allows learners to acquire phraseology that is natural, fluent, and appropriate for various contexts. Through consistent exposure to authentic usage, learners develop not only linguistic competence but also pragmatic and cultural sensitivity.

References:

1. Mc Enery T., & Hardie A. Corpus linguistics: Method, theory and practice. Cambridge University Press, 2012.
2. Sinclair J. Corpus, concordance, collocation. Oxford University Press, 1991.
3. Wray A. Formulaic language and the lexicon. Cambridge University Press, 2002.
4. Sinclair J. Trust the text: Language, corpus and discourse. – Routledge, 2004.
5. Hoey M. Lexical priming: A new theory of words and language. – Routledge.
6. Nesselhauf N. Collocations in a learner corpus. John Benjamins, 2005.
7. Rasulova M. Korpus lingvistikasi: nazariya va amaliyot asoslari. – Toshkent: Fan va texnologiya nashriyoti, 2020.
8. Bozorov F. Tilshunoslikda frazeologik birliklarni o‘rgatish metodikasi. – Samarqand: SamDU nashriyoti, 2019.
9. Juraev N. Ingliz tili o‘qitishda kollokatsiyalarni o‘rgatishning zamonaviy yondashuvlari. O‘zbekiston filologiyasi ilmiy jurnali, №3, 2021. – B. 45-52.
10. Xudoyberganova, D. Leksik birliklar va ularning korpus asosida tahlili. Filologiya masalalari, №1, 2018. – B. 63-70.