www.volsu.ru

DOI: https://doi.org/10.15688/jvolsu2.2024.4.9

UDC 81’32
LBC 81.11

Submitted: 01.03.2024
Accepted: 13.05.2024

COMBINABILITY AND STABILITY ANALYSIS OF LEXICAL UNITS BY STATISTICAL METHODS (EXEMPLIFIED BY THE VERB *TAKE*)

Marina S. Matytcina
Lipetsk State Technical University, Lipetsk, Russia

Olga N. Prokhorova
Belgorod State National Research University, Belgorod, Russia

Igor V. Chekulai
Belgorod State National Research University, Belgorod, Russia

Abstract. This article is devoted to the issues related to the definition of stable word combinability in speech. The research relevance is sustained by the existing need in profound linguistic knowledge about the factors that determine the formation of stable relationships between the elements of a word combination. The English Web Corpus (enTenTen) and its subcorpora are chosen as the source. The authors consider bigrams of a two-word combination: the verb *take* with an adjacent word. In addition to a critical examination of the measures used to determine word cohesion, the nature of the relationships between collocation elements is analysed. Particular attention is paid to the comparison of collocations in subcorpora, which contain texts of different genres and topics. More than 100 bigrams obtained through the association measures *t-score*, *MI-score* and *Log Dice* are analysed. The *t-score* measure differs across the investigated subcorpora, which demonstrates the correlation of the findings with the size of the subcorpora. It is concluded that it is not possible to determine the degree of stability of the associative relationship in the bigrams of the verb *take* based on this measure alone. The data obtained using the *MI-score* and *Log Dice* measures show little difference between subcorpora, demonstrating their independence of the corpus size. The variable nature of the relationships between the collocation elements has been revealed to lie in the dependency of the degree of coherence of words in a word combination on the frequency of their occurrence in the texts of different genres, registers and modalities. Special attention is given to the issue of identifying the degree of effectiveness of the measures in extracting verb collocations and their application to specific professional tasks.

Key words: linguistic corpus, subcorpus, collocation, measures of association, English Web Corpus (enTenTen), t-score, MI-score, Log Dice.

Citation. Matytcina M.S., Prokhorova O.N., Chekulai I.V. Combinability and Stability Analysis of Lexical Units by Statistical Methods (Exemplified by the Verb *Take*). *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], 2024, vol. 23, no. 4, pp. 106-118. (in Russian). DOI: https://doi.org/10.15688/jvolsu2.2024.4.9

УДК 81’32
ББК 81.11

Дата поступления статьи: 01.03.2024
Дата принятия статьи: 13.05.2024

ВОЗМОЖНОСТИ ИЗУЧЕНИЯ СОЧЕТАЕМОСТИ И УСТОЙЧИВОСТИ ЛЕКСИЧЕСКИХ ЕДИНИЦ СТАТИСТИЧЕСКИМИ МЕТОДАМИ (НА ПРИМЕРЕ ГЛАГОЛА *TAKE*)

Марина Станиславовна Матыцина
Липецкий государственный технический университет, г. Липецк, Россия

Ольга Николаевна Прохорова
Белгородский государственный национальный исследовательский университет, г. Белгород, Россия

Игорь Владимирович Чекулай

Белгородский государственный национальный исследовательский университет, г. Белгород, Россия

Аннотация. Статья посвящена вопросам определения устойчивой сочетаемости слов в речи с применением различных мер ассоциации на примере лингвистического корпуса. Актуальность исследования обусловлена существующей в лингвистике потребностью углубления знаний о факторах, детерминирующих формирование устойчивых отношений элементов внутри словосочетания. В качестве источника избран *English Web Corpus* (*enTenTen*) и его подкорпусы. Материалом для анализа послужили биграммы двухсловного сочетания: глагола *take* с соседним словом. Наряду с критическим рассмотрением мер, используемых для установления связности слов, описан характер отношений между элементами коллокации. Особое внимание уделено сравнению коллокаций в подкорпусах, содержащих тексты разных жанров и тематики. Проанализировано более 100 биграмм, извлеченных посредством мер ассоциации *t-score*, *MI-score* и *Log Dice.* Установлено, что показатели меры *t-score* различаются в изучаемых подкорпусах, показывают зависимость полученных данных от размера подкорпусов. Делается вывод о том, что вычисление степени устойчивости ассоциативной связи биграмм глагола *take*, основанное только на этом показателе, невозможно. Данные, полученные с помощью мер *MI-score* и *Log Dice*, свидетельствуют о незначительной разнице между подкорпусами, что демонстрирует независимость таких показателей от размера корпуса. Выявлено, что вариативный характер отношений между элементами коллокации заключается в зависимости степени связности слов в словосочетании от частоты их встречаемости в текстах разных жанров, регистров и модальности. М.С. Матыциной подготовлен общий план исследования, осуществлен сбор необходимой информации из корпуса. О.Н. Прохоровой разработана методика анализа, выполнено обобщение материала. И.В. Чекулаем интерпретированы результаты проведенной научной работы.

Ключевые слова: лингвистический корпус, подкорпус, коллокация, меры ассоциации, English Web Corpus (enTenTen), t-score, MI-score, Log Dice.

Цитирование. Матыцина М. С., Прохорова О. Н., Чекулай И. В. Возможности изучения сочетаемости и устойчивости лексических единиц статистическими методами (на примере глагола *take*) // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2024. – Т. 23, № 4. – С. 106–118. – DOI: https://doi.org/10.15688/jvolsu2.2024.4.9

Введение

В лингвистике и методике обучения иностранному языку устойчивой сочетаемости слов уделяется большое внимание, что выразилось в возникновении понятия «коллокация». Следует отметить, что соответствующий ему термин по-разному толкуется в словарях и в литературе по языкознанию.

Большинство определений отражает понимание коллокаций как сочетания двух или более слов, совместно появляющихся в тексте [Sinclair, 1991; Stubbs, 1995]. Р. Барчфилд, характеризуя лингвистическую концепцию У. Фаулера, пишет о том, что, по У. Фаулеру, термин *collocation* (коллокация) в лингвистическом аспекте впервые был употреблен Дж.Р. Фертом для обозначения совместной встречаемости отдельных слов. В работе Modern English Usage У. Фаулер приводит примеры часто встречаемых сочетаний отдельных слов и обозначает эту связность как неотъемлемую часть организации языка [Burchfield, 1996, p. 158]. The Concise Oxford Dictionary of Linguistics дает следующую трактовку коллокации: *a relation within a syntactic unit between individual lexical elements* (отношение между отдельными лексическими элементами внутри синтаксической единицы) и приводит следующий пример: *My computer hates me...* , где *computer* и *hates* находятся в отношениях коллокации к друг другу (The Concise..., 2014). И.А. Мельчук в работе «О терминах “устойчивость” и “идиоматичность”» говорит об устойчивости как о «сочетании определенных элементов, в котором эти элементы встречаются гораздо чаще, чем в других сочетаниях» [Мельчук, 1960, с. 73]. Кроме того, исследователь отмечает, что «все устойчивые сочетания относятся к несвободным, но не все несвободные сочетания являются устойчивыми (в узком смысле), то есть имеют высокую степень устойчивости» [Мельчук, 1960, с. 75].

Как показано в работе В.П. Захарова, устойчивость присуща всем сочетаниям, при

этом она варьируется от нуля до единицы. Устойчивость равна нулю, если элементы не встречаются в таком сочетании, и – единице, если оно постоянно воспроизводится в речи и один элемент предсказывает другой [Захаров, 2005]. Комплексный семасиологический и ономасиологический анализ коллокаций, проведенный Н.Л. Шамне и Л.Н. Ребриной, позволил получить информацию об их структурных, семантических и синтагматических характеристиках таких сочетаний: объектность, субъектность, внутренняя и внешняя переходность [Шамне, Ребрина, 2015]. Как считают С. Ханстон и С. Лавиоза, коллокация – это тенденция к расположению слов в тексте рядом друг с другом. Иными словами, они находятся вместе. Однако исследователи отмечают, что если два слова часто встречаются рядом, это не обязательно означает, что они имеют высокую степень связности. Например, для любого слова, по которому ведется поиск коллокатов, существует высокая вероятность того, что оно будет совпадать с некоторыми из наиболее часто встречающихся слов в английском языке, например *the*, *a* и т. д. Поэтому список коллокатов не следует принимать безоговорочно [Hunston, Laviosa, 2000]. По мнению С. Ханстона, коллокация – это тенденция к изменению значения слов при их совместном употреблении [Hunston, 2002, p. 68]. Американский лингвист Дж. Хилл отмечает распространенность коллокаций и их важность в изучении иностранного языка. Он утверждает, что коллокации могут составлять 70 % того, что мы говорим, слышим, читаем или пишем [Hill, 2000, p. 53]. Дж. Синклер исследует два принципа организации текста: принцип открытого выбора и принцип идиом [Sinclair, 1991]. Согласно принципу открытого выбора текст рассматривается как результат различных семантических комбинаций, где единственным сдерживающим фактором являются грамматические ограничения. В соответствии с данным принципом, в каждый слот может помещаться любое слово, ограниченное лишь грамматикой. Например, в грамматически правильно построенном сочетании переходного глагола и объекта любое слово может оказаться как в первом, так и во втором слоте. Принцип идиом заключается в наличии гораздо большего числа ограничений и сдерживающих факторов. В определенной степени выбор одного слова определяет выбор другого слова [Sinclair, 1991]. Владение коллокациями позволяет пользователям употреблять статистически устойчивые словосочетания подобно носителям языка и, таким образом, производить впечатление на собеседников или читателей [Henriksen, 2013]. В ряде работ показано, что использование коллокаций снижает когнитивную нагрузку и позволяет направить когнитивную энергию на более важные аспекты языка, такие, как организация дискурса и успешное взаимодействие коммуникантов (см., например: [O'Keefee, McCarthy, Carter, 2007]).

Проблематика, связанная с наличием в языке коллокаций, начала активно изучаться с развитием корпусной лингвистики, которая позволила проверить на больших массивах текстов, как слова сочетаются друг с другом. Ученые отмечают, что посредством компьютерных программ «можно решать не только относительно простые задачи, типа построения частотного анализа или синтаксического и морфологического разбора текста, но и более сложные, такие как семантический анализ, автоматическое определение стиля текста или даже его возможного автора» [Филимонов и др., 2020, с. 57]. По мнению С. Ханстона и С. Лавиоза, важно помнить, что любая информация корпуса применима только к исследуемым данным [Hunston, Laviosa, 2000]. Она не обязательно может быть применена к языку в целом. С появлением корпусной лингвистики термин *collocation* стал обозначаться как статистическая сочетаемость (см., например: [Захаров, Хохлова, 2010]).

В настоящее время статистические методы все активнее используются для проведения исследований лексической сочетаемости слов. Поскольку информация об устойчивых словосочетаниях не всегда последовательно отражается в словарях и граница между ними и свободными словосочетаниями определяется достаточно субъективно, возникла и необходимость ввести некий порог устойчивости, выше которого словосочетание можно отнести к устойчивым, а ниже – к свободным. Использование статистических методов позволяет определить такой порог на основе больших корпусов и статистических показателей, обычно называемых статистическими мерами или мерами ассоциации (далее – МА). В.П. Захаров и М.В. Хохлова в работе «Ана-

лиз эффективности статистических методов обнаружения коллокаций в текстах на русском языке» отмечают, что меры ассоциации «учитывают как частоту совместной встречаемости, так и другие параметры, прежде всего частоту в данном корпусе каждого отдельного элемента» [Захаров, Хохлова, 2010]. Использование МА при извлечении коллокаций дает возможность проанализировать не только результаты в качественном и количественном аспектах, но и дать оценку «силе притяжения» слов [Hunston, 2002; McEnery, Hardie, 2011].

При выявлении статистически устойчивых сочетаний с использованием инструментария лингвистического корпуса основным семантическим компонентом словосочетания признается изучаемое слово, а вспомогательным компонентом словосочетания, или коллокатом, значится слово, которое чаще всего сочетается с основным компонентом словосочетания в данном корпусе. Как отмечает П. Бейкер, размер диапазона контекстного окна влияет на количество выявленных коллокатов [Baker, 2006, p. 103]. Другими словами, если брать более широкий диапазон, то вероятность того, что в результаты будут включены слова, не будучи коллокатами, увеличивается. В этой связи представляется важным показать возможности корпуса в определении устойчивой сочетаемости слов, соотнести результаты, полученные на основе различных мер ассоциации, сравнить наиболее популярные меры.

Материал и методы

В данной статье предметом рассмотрения являются результаты автоматического выделения глагольных статистически устойчивых словосочетаний, в которых поисковым запросом является семантически главный компонент – глагол *take*, поскольку он образует большое количество сочетаний с разной степенью связанности. Исследование таких словосочетаний проводилось с помощью мер ассоциации *t-score*, *MI-score* и *Log Dice*, позволяющих оценить степень связности компонентов словосочетания, с целью определения их эффективности.

Материалом для проведения исследования послужили данные бесплатной версии корпуса английского языка – English Web Corpus (EWC), состоящего из интернет-текстов различного объема и содержания, который создан с использованием технологий, специализирующихся на сборе только лингвистически ценного веб-контента. Последняя версия корпуса состоит из 52 млрд слов. На рисунке 1 показан состав корпуса из наиболее часто посещаемых доменов верхнего уровня.

Информация об объеме корпуса приведена в таблице 1.

Корпус включает в себя несколько подкорпусов, которые представлены в таблице 2.

EWC содержит тексты разных жанров и тематики. Жанры письменной речи разделены на четыре группы: блоги, дискуссии, художественная литература, юридические тексты, новости, справочники и энциклопедии. Тематика текстов разнообразна: искусство, красота и мода, автомобили и велосипеды, культура и развлечения, экономика, финансы и бизнес, игры, здоровье, история, хобби, дом, семья и дети, природа и окружающая среда, домашние и дикие животные, политика и правительство, религия, наука, спорт, информационные технологии, путешествие и туризм и многое другое. Исследуемые жанры охватывают 17,5 % корпуса, то есть 10,8 млрд токенов. Темы составляют 12,2 % корпуса, то есть 7,5 млрд токенов. На рисунке 2 показана количественная представленность текстов корпуса разной тематики.

Несмотря на наличие десятков статистических мер, используемых для вычисления степени связности между коллокатами, лишь немногие из них используются для исследований. Возможно, причина заключается в присутствии только некоторых статистических мер в программных средствах извлечения коллокаций в разных корпусах. В работе мы рассмотрим степень устойчивости ассоциативной связи биграмм глагола *take*, которая может варьироваться в зависимости от лингвистических условий (например, при наличии разных социальных и ситуативных характеристик), представленных в подкорпусах EWC, используя три меры ассоциации: *t-score*, *MI-score* и *Log Dice*. С этой целью, а также с целью проиллюстрировать влияние различных жанров, регистров и модальности на степень связности компонентов словосочетаний, были выбраны четыре подкорпуса, составляющих наибольший процент от объема всего корпуса. Исследуемые подкорпусы представлены в таблице 3.

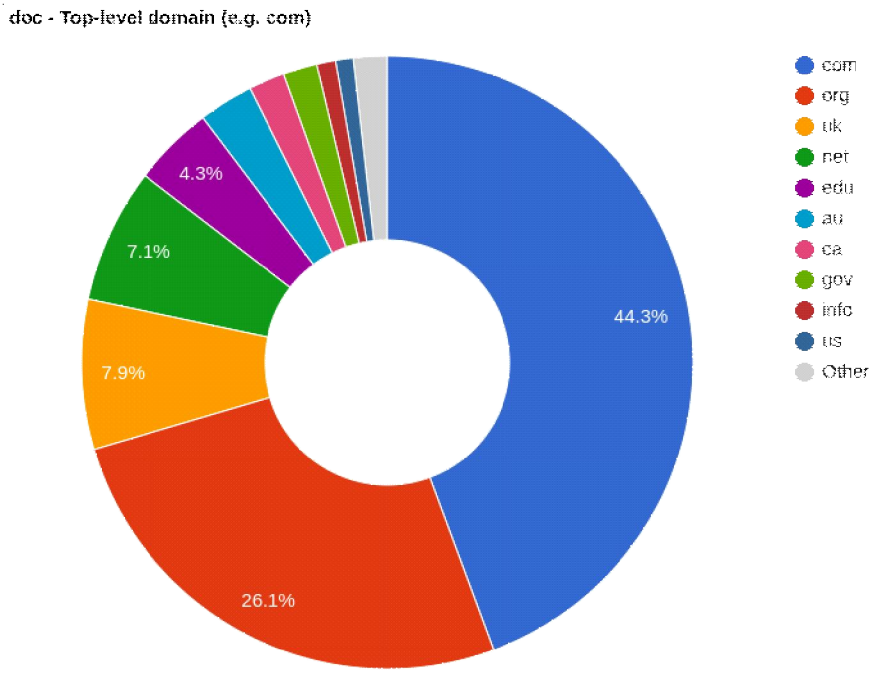


Рис. 1. Домены в составе EWC
Fig. 1. Domains within the EWC

Таблица 1. **Количественная характеристика EWC**

Table 1. **Quantitative characteristics of the EWC**

Единицы корпуса	Количество единиц
Токены	61,585,997,113
Слова	52,268,286,493
Предложения	2,852,972,274
Веб-страницы	120,252,162

Таблица 2. **Список подкорпусов, имеющихся в EWC**

Table 2. **List of subcorpora available in the EWC**

Подкорпус	Токены	% в корпусе
Australian domain.au	1,083,884,536	1.76
Canadian domain.ca	1,279,711,284	2.078
EU domain.eu	178,200,834	0.289
English Wikipedia	2,781,502,596	4.516
Genre Blog	1,569,499,442	2.548
Genre Discussion	2,103,533,595	0.726
Genre Fiction	1,030,493,934	1.673
Genre Legal	652,370,863	1.059
Genre News	2,420,719,017	3.931
Genre Reference/Encyclopedia	3,047,342,438	4.948
Indian domain.in	275,247,190	0.447
Irish domain.ie	343,876,212	0.558
New Zealand domain.nz	318,843,917	0.518
Topic Arts	191,659,187	0.311
Topic Beauty & Fashion	54,111,137	3.06
Topic Cars & Bikes	293,808,521	0.477
Topic Culture & Entertainment	989,990,028	1.607

Окончание таблицы 2

End of Table 2

Подкорпус	Токены	% в корпусе
Topic Economy, Finance & Business	511,642,058	0.831
Topic Education	235,276,797	0.382
Topic Games	398,193,384	0.647
Topic Health	480,090,118	0.78
Topic Hobbies	103,913,047	0.169
Topic Nature & Environment	1,034,647,808	2.399
Topic Pets & Animals	1,566,713,474	3.633
UK domain.uk	3,466,969,061	5.629
US domain.us	440,106,116	0.715

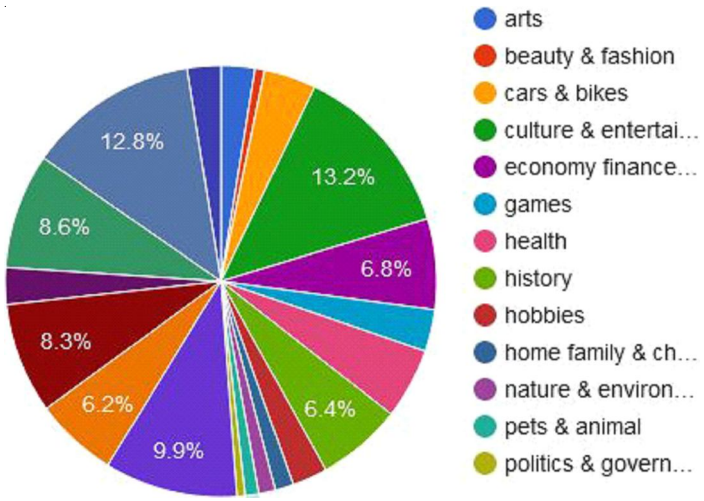


Рис. 2. Тематика текстов EWC

Fig. 2. Subject matter of the texts of the EWC

Таблица 3. **Исследуемые подкорпусы EWC**

Table 3. **Studied subcorpora of the EWC**

Подкорпус	Токены	% в корпусе
UK domain.uk	3,466,969,061	5.629
Genre Reference/Encyclopedia	3,047,342,438	4.948
English Wikipedia	2,781,502,596	4.516
Genre News	2,420,719,017	3.931

МА *T-score*. Лингвисты определяют *t-score* по-разному: например, как показатель «надежности коллокации» [Hunston, 2002, p. 73] или как «силу связи» между коллокатами, которая «проверяет нулевую гипотезу» [Wolter, Gyllstad, 2011, p. 436]. *T-score* характеризует «статистически значимые» коллокаты, то есть те, которые появляются чаще, чем случайно. В исследовании использованы формулы расчета этих мер, предложенные В.П. Захаровым и М.В. Хохловой [Захаров, Хохлова, 2010]:

$$t\text{-}score = \frac{f(n,c) - \frac{f(n) \times f(c)}{N}}{\sqrt{f(n,c)}}.$$

С. Ханстон отмечает, что данная мера учитывает наличие коллокации во всем корпусе и используется, когда нужно доказать ее существование как результат чего-то большего, нежели «причуды» конкретного корпуса [Hunston, 2002, p. 72]. Лингвисты отмечают, что мера *t-score* выделяет частотные сочетания слов [Durrant, Schmitt, 2009; Hunston, 2002; Siyanova, Schmitt, 2008].

Очевидным способом анализа распространенности таких словосочетаний, по мнению ряда исследователей, значится вычисление абсолютной частоты, то есть фактического количества их употребления в тексте [Durrant, Schmitt, 2009, p. 167]. Несмотря на проведенную параллель между словосочетаниями, обнаруженными с помощью *t-score*, и абсолютной частотой, их нельзя рассматривать как равноценные. Показатель *t-score* дает четкое представление о коллокациях, имеющих устойчивую ассоциативную связь и появляющихся в корпусе чаще, чем случайно. Однако С. Ханстон и С. Лавиоза отмечают, что *t-score* показывает только те слова, которые важны для ключевого слова, а не те, для которых ключевое слово является важным [Hunston, Laviosa, 2000]. Согласно С. Ханстону, показатель *t-score*, равный 2 или выше, следует считать важным [Hunston, 2002, p. 72].

МА *MI-score*. Наряду с мерой *t-score* широкое распространение в корпусной лингвистике получила мера *MI-score*. Исследователи отмечают, что мера *MI* указывает на вероятность совпадения двух случайных слов, сравнивает то, что есть с тем, что могло бы быть. Однако распределение слов в языке никогда не бывает случайным, поэтому нельзя получить «ожидаемый» результат. *MI* используется как мера, показывающая, в какой степени слово «обладает информацией» о другом слове. *MI*, равный 3 или выше, допустимо интерпретировать как свидетельство того, что сочетание двух слов является коллокацией [Hunston, 2002]. Однако, по мнению П. Бейкера, один из недостатков *MI* заключается в том, что данная мера склонна придавать большое значение словам, которые редко встречаются в тексте, поэтому дает несколько искаженные результаты [Baker, 2006, p. 102]. Тем не менее *MI* полезна при вычислении степени связности тех компонентов словосочетания, у которых вероятность совместной встречаемости очень высока даже при относительно небольшом количестве случаев употребления всего сочетания.

$$MI = \log_2 \frac{f(n,c) \times N}{f(n) \times f(c)}.$$

МА *Log Dice*. *Log Dice* – это мера, которая основана на *MI*, но в отличии от *MI*, она не придает особого веса низкочастотным словосочетаниям. Ее можно рассматривать как промежуточную [Evert, 2008; Smadja, McKeown, Hatzivassiloglou, 1996].

Результаты и обсуждение

С использованием инструментария лингвистического корпуса были составлены таблицы мер по подкорпусам. Как видно на рисунках 3–6, мера *t-score* ожидаемо различается в исследуемых подкорпусах, что говорит о связи между полученными данными и размерами подкорпусов. Следовательно, сравнение подкорпусов, основанное лишь на этом показателе, невозможно, поскольку *t-score* напрямую зависит от размера корпуса, невзирая на силу связи между элементами словосочетания. При этом данные, приведенные на рисунках 3 и 4 отражают незначительные различия в значениях меры *t-score* между подкорпусами *Genre Reference*/*Encyclopedia* и *English Wikipedia*, но их трудно интерпретировать из-за отсутствия стандартизированной шкалы.

Данные, полученные с помощью мер *MI* и *Log Dice* (см. рис. 5, 6), свидетельствуют о незначительной разнице между подкорпусами, что демонстрирует их независимость от размера корпуса.

Различия в степени связности компонентов коллокации в подкорпусах EWC позволяют говорить о том, что такие лексические единицы ограничены жанрами, регистрами и модальностью текстов. Это характерно для любой коллокации, однако в некоторых случаях это особенно очевидно. Так, в сочетании *take advantage* разница в значениях меры *MI* в подкорпусах относительно невелика в отличие от *take place*, где *MI* колеблется от 11,51 до 13,07, а значения *Log Dice* – от 10,93 до 11,63. Стоит отметить и отличие в ранжировании сочетаний по степени связности компонентов в исследуемых подкорпусах. Так, словосочетание *take advantage*, как показано на рисунке 7, в самом корпусе имеет показатель *MI* выше, чем *take place*: 8.03 и 6.85, однако в подкорпусе *Genre Reference*/*Encyclopedia* (см. рис. 3) значения этого показателя практически одинаковые 12.96 и 12.94.

Collocations

	Word	Cooccurrences	Candidates	T-score	MI	LogDice
1	place	363,682	1,006,858	602.98	12.94	11.60
2	over	202,969	2,342,020	450.28	10.88	10.33
3	part	126,719	1,790,813	355.75	10.59	9.81
4	up	85,865	2,260,673	292.67	9.69	9.11
5	advantage	39,610	108,611	199.00	12.96	8.79
6	care	31,234	224,058	176.67	11.57	8.39
7	off	34,777	808,483	186.29	9.87	8.29
8	control	32,596	754,844	180.35	9.88	8.22
9	account	23,224	200,938	152.33	11.30	7.97
10	on	167,155	20,093,347	406.59	7.50	7.90

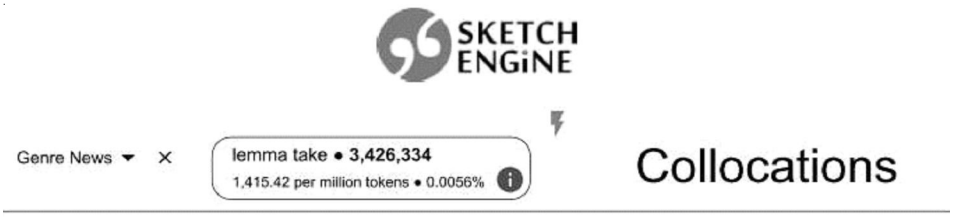
Рис. 3. Данные подкорпуса *Genre Reference/Encyclopedia*

Fig. 3. Data of the subcorpus *Genre Reference/Encyclopedia*

	Word	Cooccurrences	Candidates	T-score	MI	LogDice
1	place	346,017	933,318	588.16	13.07	11.63
2	over	198,589	2,173,107	445.42	11.05	10.40
3	part	124,346	1,641,277	352.43	10.78	9.89
4	up	80,008	2,091,242	282.54	9.80	9.11
5	advantage	36,647	92,926	191.41	13.16	8.77
6	care	28,990	198,174	170.21	11.73	8.38
7	control	31,660	471,198	177.82	10.61	8.38
8	off	33,343	750,709	182.42	10.01	8.33
9	on	160,048	18,834,555	398.03	7.63	7.93
10	into	36,544	2,860,444	190.52	8.21	7.76

Рис. 4. Данные подкорпуса *English Wikipedia*

Fig. 4. Data of the subcorpus *English Wikipedia*

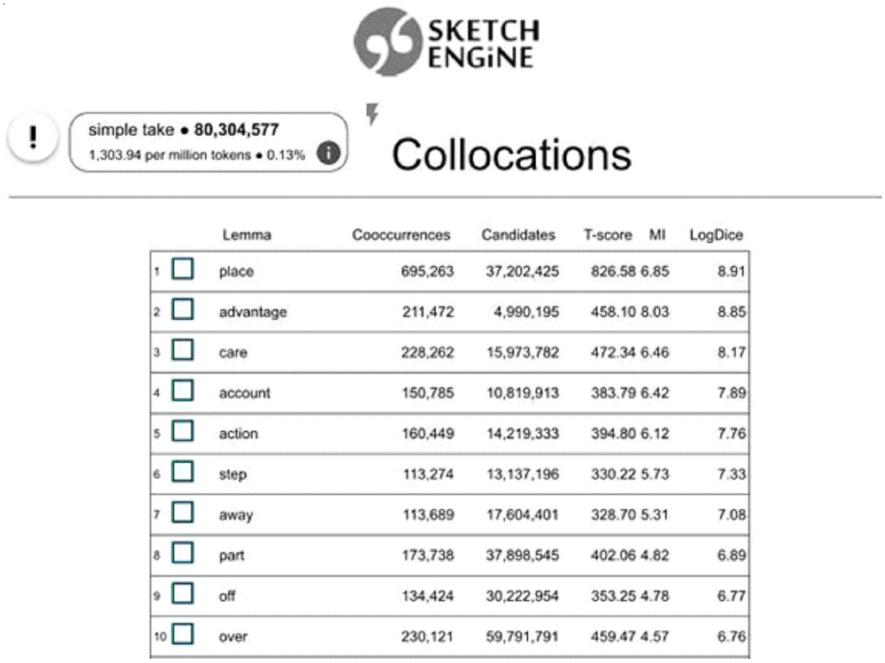


	Word	Cooccurrences	Candidates	T-score	MI	LogDice
1	place	259,233	924,701	509.05	12.30	10.93
2	over	120,353	2,833,278	346.46	9.58	9.30
3	advantage	68,184	172,858	261.08	12.79	9.28
4	part	70,785	1,118,607	265.82	10.15	9.00
5	care	56,193	559,449	236.92	10.82	8.85
6	action	50,541	437,790	224.70	11.02	8.74
7	away	44,690	662,659	211.23	10.24	8.48
8	up	73,185	3,620,635	269.78	8.51	8.41
9	into	62,780	2,725,588	249.95	8.69	8.39
10	account	36,490	244,113	190.95	11.39	8.35

Рис. 5. Данные подкорпуса *Genre News*
Fig. 5. Data of the subcorpus *Genre News*

	Word	Cooccurrences	Candidates	T-score	MI	LogDice
1	place	418,202	1,725,072	646.46	11.51	10.97
2	part	189,109	1,963,740	434.49	10.18	9.77
3	account	91,599	476,542	302.52	11.18	9.07
4	over	134,467	3,502,365	365.90	8.85	9.00
5	advantage	82,779	197,449	287.66	12.30	9.00
6	care	77,883	1,055,669	278.76	9.79	8.69
7	up	130,446	5,501,622	359.91	8.16	8.65
8	into	110,778	4,112,407	331.81	8.34	8.62
9	look	70,626	1,317,260	265.34	9.33	8.49
10	action	55,890	569,572	236.21	10.21	8.33

Рис. 6. Данные подкорпуса *UK domain.uk*
Fig. 6. Data of the subcorpus *UK domain.uk*



	Lemma	Cooccurrences	Candidates	T-score	MI	LogDice
1	place	695,263	37,202,425	826.58	6.85	8.91
2	advantage	211,472	4,990,195	458.10	8.03	8.85
3	care	228,262	15,973,782	472.34	6.46	8.17
4	account	150,785	10,819,913	383.79	6.42	7.89
5	action	160,449	14,219,333	394.80	6.12	7.76
6	step	113,274	13,137,196	330.22	5.73	7.33
7	away	113,689	17,604,401	328.70	5.31	7.08
8	part	173,738	37,898,545	402.06	4.82	6.89
9	off	134,424	30,222,954	353.25	4.78	6.77
10	over	230,121	59,791,791	459.47	4.57	6.76

Рис. 7. Данные EWC

Fig. 7. Data of the EWC

Если сравнить три подкорпуса: *Genre Reference*/*Encyclopedia*, *English Wikipedia*, *Genre News*, содержащие энциклопедические статьи (см. рис. 3–5), то разница в показателях *MI* и *Log Dice* незначительна, хотя их значения различаются практически на единицу (*Log Dice* для *take advantage* имеет диапазон от 8,77 до 9,28; *take place* от 10,93 до 11.60). Эти результаты показывают, что отношения между элементами коллокации и стилем текста требуют пристального внимания. Наличие разных показателей *MI* и *Log Dice* в корпусе, представляющих один и тот же тип дискурса (три подкорпуса, содержащие энциклопедические статьи), свидетельствуют о том, что детальная категоризация полученных значений с помощью МА может быть не совсем корректной. Необходимо проведение дальнейших исследований в этой области.

Вариативный характер отношений между элементами коллокации, который заключается в зависимости степени связности от частоты их встречаемости в текстах разных жанров, регистров и модальности, имеет важное значение в определении коллокатов.

Заключение

Данное исследование представляет собой попытку рассмотреть вопросы, связанные с определением устойчивой сочетаемости слов в речи с использованием различных мер ассоциации на примере лингвистического корпуса. На данном этапе развития корпусной лингвистики выделить все факторы определения коллокаций не представляется возможным. Однако их критическое рассмотрение, установление сочетаемости слов с учетом влияния жанров, регистров и модальности текстов на характер отношений между словами в составе словосочетания представляется весьма важным в решении профессиональных задач, в частности при обучении иностранному языку. Результаты сравнительного анализа мер ассоциации *t-score*, *MI-score* и *Log Dice* EWC и его подкорпусов на примере глагола *take* показывают, что вычисление степени устойчивости ассоциативной связи биграмм глагола *take*, основанное только на показателе *t-score*, невозможно, поскольку он напрямую зависит от размера корпуса. Данные, полученные с помощью мер *MI-score* и *Log Dice*, свидетельствуют о незначительной разнице между подкорпусами, что демонстрирует их независимость от размера корпуса. Очевидно, что изучение коллокаций требует математического и лингвистического обоснования каждой МА с тем, чтобы осмысленно применять данные меры и правильно интерпретировать полученные результаты.

Несмотря на большое количество корпусных исследований сочетаемости, по-прежнему существует потребность в более глубоком понимании факторов, играющих важную роль в установлении того, что можно считать коллокациями.

СПИСОК ЛИТЕРАТУРЫ

Захаров В. П., Хохлова М. В., 2010. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии : тр. Междунар. конф. «Диалог-2010» (Бекасово, 26–30 мая 2010 г.). М. : Изд-во РГГУ. Вып. 9 (16). С. 137–143.

Захаров В. П., 2005. Корпусная лингвистика. СПб. : Изд-во СПбГУ. 48 с.

Мельчук И. А., 1960. О терминах «устойчивость» и «идиоматичность» // Вопросы языкознания. № 4. С. 73–80.

Филимонов Д. Ю., Светлов А. В., Горбань О. А., Косова М. В., Шептухина Е. М., 2020. Автоматизация процесса метаразметки архивных документов // Математическая физика и компьютерное моделирование. Т. 23, № 4. С. 56–68.

Шамне Н. Л., Ребрина Л. Н., 2015. Глагольные коллокации памяти в германских СМИ // В мире научных открытий. № 7–8 (67). С. 3097–3108.

Baker P., 2006. Using Corpora in Discourse Analysis. L. : Bloomsbury Academic. 280 p.

Burchfield R. W., 1996. The New Fowler's Modern English Usage. Oxford : Oxford University Press. 864 p.

Durrant P., Schmitt N., 2009. To What Extent Do Native and Non-Native Writers Make Use of Collocations? // IRAL-International Review of Applied Linguistics in Language Teaching. Vol. 47, iss. 2. P. 157–177. DOI:10.1515/iral.2009.007

Evert S., 2008. Corpora and Collocations // Corpus Linguistics: An International Handbook. Berlin : Mouton de Gruyter. P. 1212–1248. DOI:10.1515/9783110213881.2.1212

Henriksen B., 2013. Research on L2 Learners' Collocational Competence and Development – A Progress Report // L2 Vocabulary Acquisition, Knowledge and Use New Perspectives on Assessment and Corpus Analysis. [S.l.] : [s.n.]. P. 29–56.

Hill J., 2000. Revising Priorities: From Grammatical Failure to Collocational Success // Teaching Collocation: Further Developments in the Lexical Approach. Hove : LTP. P. 47–67.

Hunston S., 2002. Corpora in Applied Linguistics. Cambridge : Cambridge University Press. 241 p. DOI:10.1017/CBO9781139524773

Hunston S., Laviosa S., 2000. Corpus Linguistics. Birmingham : School of English : CELS. 146 p.

O'Keeffe, A., McCarthy M., Carter R., 2007. From Corpus to Classroom: Language Use and Language Teaching. Cambridge : Cambridge University Press. 332 p.

McEnery T., Hardie A., 2011. Corpus Linguistics: Method, Theory and Practice. Cambridge : Cambridge University Press. 312 p.

Sinclair J., 1991. Corpus, Concordance, Collocation. Oxford : Oxford University Press. 179 p.

Siyanova A., Schmitt N., 2008. L2 Learner Production and Processing of Collocation: A Multi-Study Perspective // Canadian Modern Language Review. Vol. 64, № 3. P. 429–458. DOI: 10.3138/cmlr.64.3.429

Smadja F., McKeown K. R., Hatzivassiloglou V., 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach // Computational Linguistics. Vol. 22, № 1. P. 1–38.

Stubbs M., 1995. Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies // Functions of Language. Vol. 2, iss. 1. P. 23–55. DOI: 10.1075/fol.2.1.03stu

Wolter B., Gyllstad H., 2011. Collocational Links in the L2 Mental Lexicon and the Inuence of L1 Intralexical Knowledge // Applied Linguistics. Vol. 32, iss. 4. P. 430–449. DOI: 10.1093/applin/amr011

ИСТОЧНИК

EWC – English Web Corpus (enTenTen). URL: https://www.sketchengine.eu/ententen-english-corpus/ (дата обращения: 20.12.2021)

СЛОВАРЬ

The Concise Oxford Dictionary of Linguistics. Oxford : Oxford University Press, 2014. 443 p.

REFERENCES

Zakharov V.P., Khokhlova M.V., 2010. Analiz effektivnosti statisticheskikh metodov vyyavleniya kollokatsiy v tekstakh na russkom yazyke [Analysis of the Effectiveness of Statistical Methods for Identifying Collocations in Russian Texts]. *Kompyuternaya lingvistika i intellektualnye tekhnologii: tr. Mezhdunar. konf. «Dialog-2010» (Bekasovo, 26–30 maya 2010 g.)* [Computer Linguistics and Intelligent Technologies. Proceedings of the International Conference "Dialogue-2010" (Bekasovo, May 26–30, 2010)]. Moscow, Izd-vo RGGU, iss. 9 (16), pp.137-143.

Zakharov V.P., 2005. *Korpusnaya lingvistika* [Corpus Linguistics]. Saint Petersburg, Izd-vo SPbGU. 48 p.

Melchuk I.A., 1960. O terminakh «ustoychivost» i «idiomatichnost» [On Terms "Sustainability" and "Idiomaticity"]. *Voprosy yazykoznaniya* [Topics in the Study of Language], no. 4, pp. 73-80.

Filimonov D.Yu., Svetlov A.V., Gorban O.A., Kosova M.V., Sheptukhina E.M., 2020. Avtomatizatsiya protsessa metarazmetki arkhivnykh dokumentov [Automation of the Process of Meta-Labelling of Archival Documents]. *Matematicheskaya fizika i kompyuternoe modelirovanie* [Mathematical Physics and Computer Simulation], vol. 23, no. 4, pp. 56-68.

Schamne N.L., Rebrina L.N., 2015. Glagolnye kollokatsii pamyati v germanskikh SMI [Verb Collocations of Memory in German Media]. *V mire nauchnykh otkrytiy* [In the World of Scientific Discoveries], no. 7-8 (67), pp. 3097-3108.

Baker P., 2006. *Using Corpora in Discourse Analysis*. London, Bloomsbury Academic. 280 p.

Burchfield R.W., 1996. The New Fowler's Modern English Usage. Oxford, Oxford University Press. 864 p.

Durrant P., Schmitt N., 2009. To What Extent Do Native and Non-Native Writers Make Use of Collocations? *IRAL-International Review of Applied Linguistics in Language Teaching*, vol. 47, iss. 2, pp. 157-177. DOI: 10.1515/iral.2009.007

Evert S., 2008. Corpora and Collocations. *Corpus Linguistics: An International Handbook*. Berlin, Mouton de Gruyter, pp. 1212-1248. DOI: 10.1515/9783110213881.2.1212

Henriksen B., 2013. Research on L2 Learners' Collocational Competence and Development – A Progress Report. *L2 Vocabulary Acquisition, Knowledge and Use New Perspectives on Assessment and Corpus Analysis*. S. l., s. n. P. 29-56.

Hill J., 2000. Revising Priorities: From Grammatical Failure to Collocational Success. *Teaching Collocation: Further Developments in the Lexical Approach*. Hove, LTP, pp. 47-67.

Hunston S., 2002. *Corpora in Applied Linguistics*. Cambridge, Cambridge University Press. 241 p. DOI: 10.1017/CBO9781139524773

Hunston S., Laviosa S., 2000. *Corpus Linguistics*. Birmingham, School of English, CELS. 146 p.

O'Keeffe A., McCarthy M., Carter R. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge, Cambridge University Press. 332 p.

McEnery T., Hardie A., 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge, Cambridge University Press. 312 p.

Sinclair J., 1991. *Corpus, Concordance, Collocation*. Oxford, Oxford University Press. 179 p.

Siyanova A., Schmitt N., 2008. L2 Learner Production and Processing of Collocation: A Multi-Study Perspective. *Canadian Modern Language Review*, vol. 64, no. 3, pp. 429-458. DOI: 10.3138/cmlr.64.3.429

Smadja F., McKeown K.R., Hatzivassiloglou V., 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, vol. 22, no. 1, pp. 1-38.

Stubbs M., 1995. Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies. *Functions of Language*, vol. 2, iss. 1, pp. 23-55. DOI: 10.1075/fol.2.1.03stu

Wolter B., Gyllstad H., 2011. Collocational Links in the L2 Mental Lexicon and the Inuence of L1 Intralexical Knowledge. *Applied Linguistics*, vol. 32, iss. 4, pp. 430-449. DOI: 10.1093/applin/amr011

SOURCE

English Web Corpus (enTenTen). URL: https://www.sketchengine.eu/ententen-english-corpus/ (accessed Dec. 20, 2021)

DICTIONARY

The Concise Oxford Dictionary of Linguistics. Oxford, Oxford University Press, 2014. 443 p.

Information About the Authors

Marina S. Matytcina, Doctor of Sciences (Philology), Professor, Department of Foreign Languages, Lipetsk State Technical University, Moskovskaya St, 30, 398055 Lipetsk, Russia, lipmarina@gmail.com, https://orcid.org/0000-0001-6102-4397

Olga N. Prokhorova, Doctor of Sciences (Philology), Professor, Director, Institute of Intercultural Communication and International Relations, Belgorod State National Research University, Pobedy St, 85, Bld. 10, 308015 Belgorod, Russia, prokhorova@bsu.edu.ru, https://orcid.org/0000-0001-9441-819X

Igor V. Chekulai, Doctor of Sciences (Philology), Professor, Department of English Philology and Intercultural Communication, Belgorod State National Research University, Pobedy St, 85, Bld. 10, 308015 Belgorod, Russia, chekulai@bsu.edu.ru, https://orcid.org/0000-0001-8599-1699

Информация об авторах

Марина Станиславовна Матыцина, доктор филологических наук, профессор кафедры иностранных языков, Липецкий государственный технический университет, ул. Московская, 30, 398055 г. Липецк, Россия, lipmarina@gmail.com, https://orcid.org/0000-0001-6102-4397

Ольга Николаевна Прохорова, доктор филологических наук, профессор, директор Института межкультурной коммуникации и международных отношений, Белгородский государственный национальный исследовательский университет, ул. Победы, 85, корп. 10, 308015 г. Белгород, Россия, prokhorova@bsu.edu.ru, https://orcid.org/0000-0001-9441-819X

Игорь Владимирович Чекулай, доктор филологических наук, профессор кафедры английской филологии и межкультурной коммуникации, Белгородский государственный национальный исследовательский университет, ул. Победы, 85, корп. 10, 308015 г. Белгород, Россия, chekulai@bsu.edu.ru, https://orcid.org/0000-0001-8599-1699