УДК 159.946.3

ИЗУЧЕНИЕ КОЛЛОКАЦИЙ НА ОСНОВЕ ЛИНГВИСТИЧЕСКИХ КОРПУСОВ ТЕКСТОВ

© Татьяна Юрьевна ПАВЕЛЬЕВА
Липецкий государственный педагогический университет,
г. Липецк, Российская Федерация, кандидат педагогических наук,
доцент кафедры лингвистики и межкультурной коммуникации,
e-mail: tatianapavelieva@live.com

Рассмотрена проблема выявления и изучения коллокаций на основе лингвистических корпусов текстов. Рассмотрены основные понятия корпусной лингвистики, такие как репрезентативность и объем корпуса, корпусный менеджер (конкорданс), особенности корпусов текстов. Приведены определения коллокаций с точки зрения традиционной и корпусной лингвистики. Описаны различные подходы к определению контекста, как необходимого условия для выделения коллокаций. Особенно подчеркнуто, что в корпусной лингвистике коллокации рассматриваются как статистически устойчивые словосочетания. Проанализированы основные преимущества лингвистических корпусов текстов для лингвистических исследований, такие как полнота и релевантность текстовой информации и некоторые статистические методы обработки языка, в частности мера MI, или коэффициент взаимной информации, которая сравнивает зависимые контекстно-связанные частоты с независимыми и которая позволяет выделять наиболее редкие коллокации. Приведены примеры выявления и изучения коллокаций лексической единицы time на основе данных Британского национального корпуса.

Ключевые слова: коллокация; корпусная лингвистика; корпус текстов; контекст; коэффициент взаимной встречаемости.

DOI: 10.20310/1810-0201-2016-21-3/4(155/156)-56-61

В процессе проникновения современных информационно-коммуникационных технологий во все сферы науки, и в частности, в филологической науки, возрастает популярность использования лингвистических корпусов текстов в исследовании различных аспектов языка. За последние несколько лет в методической литературе появился целый цикл методических исследований, посвященный обучению учащихся и студентов лексической и грамматической сторонам иноязычной речи посредством разных видов лингвистических корпусов [1–4]. Анализ этих и других исследований свидетельствует о том, что авторы пришли к определенному согласию относительно понятийного содержания термина «лингвистический корпус». Под ним принято понимать организованный массив отобранных и размеченных по определенной методике и представленном в электронном виде текстов.

В.В. Рыков считает, что «корпус текстов может рассматриваться как достаточно сложно организованная антология речевой деятельности, отражающая в себе все жанровое разнообразие представленного в нем рода словесности и занимает промежуточное положение между реальными коммуникативными процессами в обществе, которые он представляет, и формализованной лингвистической теорией, для которой он является источником для исследования» [5].

Одной из важнейших характеристик корпуса текстов является репрезентативность. Так как корпус – это своеобразная модель языка, его репрезентативность определяет достоверность полученных на его основе данных, поэтому вопрос может рассматривать как «проблему адекватного отражения, адаптации или интеграции больших массивов текстов или некоторых иных фрагментов речевой деятельности в существенно меньший по объему корпус текстов» [6]. В корпусной лингвистике под репрезентативностью понимается сбалансированное и пропорциональное представление текстов в корпусе. Наиболее репрезентативными являются: Национальный корпус английского языка, Корпус современного английского языка (американский вариант) и ряд других. Обширный лингвистический материал корпусов позволяет исследователям решать различные виды лингвистических задач, в частности, изучение коллокаций. В этом смысле лингвистический корпус представляет неограниченные возможности, т. к. в процессе обра-

ботки текста происходит максимальная опора на контекст, т. е. важным преимуществом корпуса текстов является то, что языковые данные разного типа находятся в корпусе в своей естественной контекстной форме. Причем в рамках корпусных исследований контекст может рассматриваться в разных смыслах. Так Е.В. Ягунова выделяет несколько уровней понимания контекста:

1) минимальный контекст, в котором реализуются лексические и морфолого-синтаксические явления;

2) текстовый контекст, включающий в себя фрагменты текста вплоть до текста целиком;

3) контекст, предполагающий учет текстов определенного типа (заданного функционального стиля, отобраной коллекции текстов и т. д.) [7].

Именно лингвистический корпус текстов дает возможность изучать лексическую сочетаемость, или явление коллокаций в контексте.

Термин коллокация занимает одно из ведущих мест в современной лингвистике. В широком смысле коллокация – это сочетание двух или более слов, имеющих тенденцию к совместной встречаемости. В современной лингвистике существуют различные определения понятия «коллокация». Большинство определений строится на явлении семантико-грамматической взаимообусловленности элементов словосочетания [7].

Термин «коллокация» впервые был употреблен У. Фаулером в фундаментальном труде “English Grammar” (1857) при изучении внутренней структуры группы слов. В русскоязычной научной литературе он появился в словаре лингвистических терминов О.С. Ахмановой, которая определяет коллокации как «лексико-фразеологическую обусловленную сочетаемость слов в речи для реализации их полисемии» [8]. Е.Г. Борисова признает коллокацией несвободное неидиоматическое словосочетание, в котором свободный компонент не меняет своего смысла, а смысл несвободного отличается от того, который реализуется в свободных словосочетаниях [9].

Мы опираемся на определение коллокации, предложенное А.Н. Барановым и Д.О. Добровольским: «это слабоидиоматичные фразеологизмы преимущественно со структурой словосочетания, в которых семантически главный компонент (база) употреблен в своем прямом значении, а сочетаемость со вспомогательным компонентом (коллокатором) может быть задана в терминах семантического класса, но выбор конкретного слова предопределен узусом» [10]. Слово, которое сохраняет свое значение, называется ключевым, или свободным компонентом: слово *to make impact on*, слово scenery в сочетании beautiful scenery. Слово, выбор которого определяется традицией, зависит от ключевого компонента, называется несвободным компонентом. Все несвободное сочетание должно храниться в памяти (в словаре) [11]. И.А. Мельчук утверждает, что «под сочетаемостью языковой единицы понимается число других элементов, с каждым из которых данный элемент может вступать в определенное отношение» [12].

В настоящее время термин коллокация широко используется в корпусной лингвистике, где понятие коллокации содержательно упрощается по сравнению с традиционной лингвистикой. Такой подход исследователи называют статистическим [3]. Коллокации в корпусной лингвистике определяют как статистически устойчивые словосочетания. При этом такое статистически устойчивое сочетание может быть как фразеологизированным, так и свободным. Необходимо отметить, что коллокация не идентична фразеологизму, т. к. под последним традиционно понимается идиома. Коллокации же обычно образуются в процессе изменения значений слов в результате переосмысления отдельных компонентов сочетания. Главная особенность коллокации – это сохранение лексического значения одним из ее компонентов при устойчивости словосочетания в целом и отсутствие или ослабление значения образности и эмоционально оценочной нагрузки.

Важность изучения коллекций не вызывает сомнения, т. к. они присутствуют во всех языках и представляют главную особенность языка. «Коллокация уточняет и конкретизирует значение слов, которые ее образуют, точное значение в любом контексте обусловлено окружением данного слова – его распространителем, или коллокатами, которые объединены вокруг него и образуют коллокацию. Объединение слов в коллокацию является фундаментальным для всего языкового использования» [2].

Как уже было отмечено ранее, корпус текстов представляет неограниченные возможности для изучения и выявления коллокаций. Большие национальные корпуса, такие как постоянно пополняющийся Банк английского языка или Британский национальный корпус составляются по возможности максимально большими с целью отражения состояния языка. Что сколько-нибудь значащие статистические данные и результаты можно получать только на достаточно большом объеме материала. В особенности это утверждение справедливо в отношении выявления коллокаций. Для изучения коллокаций эффективным инструментом является конкорданс. Конкорданс – это строки текста заданной длины, показывающие использование слова в контексте. Кроме того, в корпусной лингвистике существует несколько способов для вычисления степени связности частей коллокации. Одной из таких мер считается MI, т. е. коэффициент взаимной зависимости. Этот показатель отражает т. н. притяжение между двумя словами, или вероятность того, что если одно слово появляется в строке, слово, составляющее с ним коллокацию, появится в ближайшей окрестности искомого слова. Корпус Банк английского языка и Британский национальный корпус располагает инструментами подсчета показателя взаимной информации (MI), который может быть как положительным, так и отрицательным. Чем выше показатель, тем сильнее взаимное притяжение слов. Если значение MI > 1, то данное сочетание является статистически значимым, а если примерно равно 0, то слова встречаются в паре крайне редко и могут считаться случайными сочетанием [13].

В нашем случае целью являлось изучить коллекции слова time на основе Британского национального корпуса и распределить коллокации по степени устойчивости. Если мера MI > 3, коллокация считается высокоустойчивой, если > 1, то просто устойчивой, если ~ 0, то случайной. Представленные в корпусе данные располагаются в порядке убывания частотности употребления.

Анализ данных корпуса (рис. 1) показывает, что наиболее часто употребительными являются сочетания существительного time с порядковым числительным first, а также прилагательными same, some, long и др. В то же время к наиболее устойчивым сочетаниям можно отнести сочетания same time, first time, long time, every time, т. к. их коэффициент взаимной встречаемости значительно выше.

На рис. 2–3 представлена сочетаемость лексической единицы time с отдельными частями речи, в частности с глаголами и прилагательными. Данные поиска в корпусе свидетельствуют о том, что:

1) в сочетании с глаголами (time + verb) данная лексическая единица имеет высокоустойчивые сочетания: **spent time, spend time, time consuming, time wasting, time allocated**;

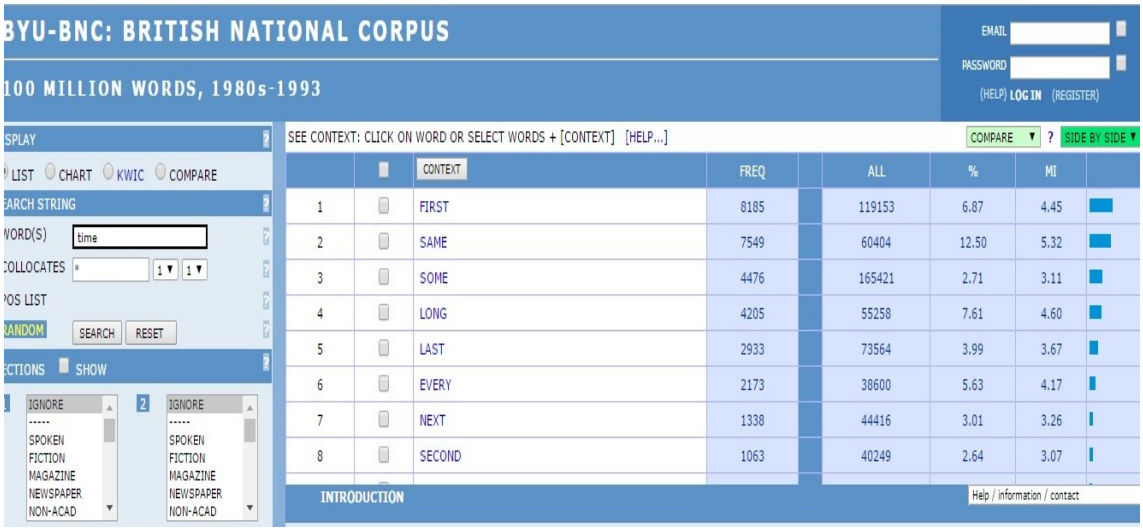

Рис. 1. Данные корпуса о сочетаемости лексической единицы time

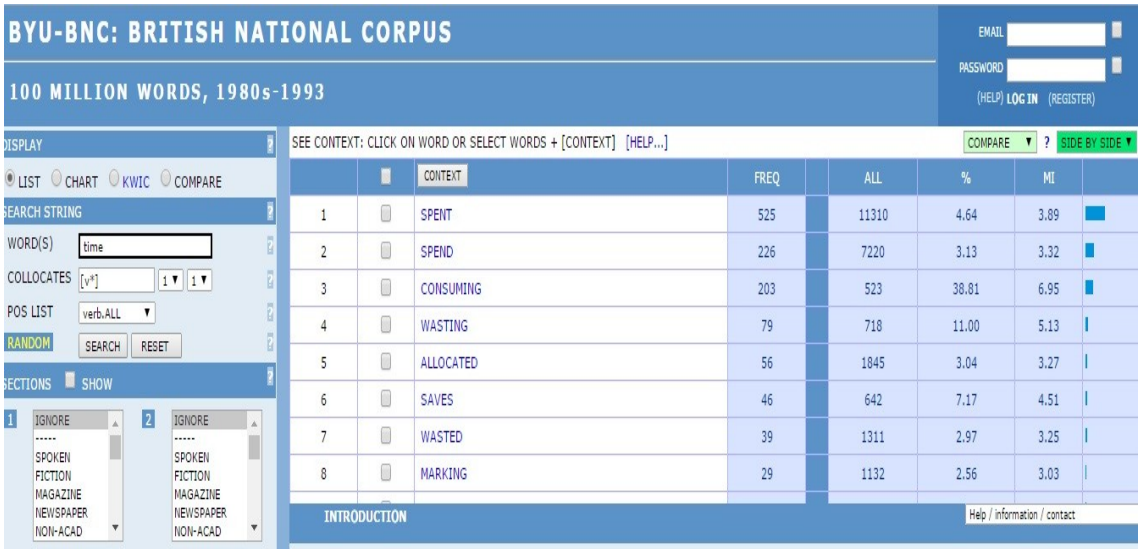

Рис. 2. Данные корпуса о сочетаемости лексической единицы time по модели time + verb

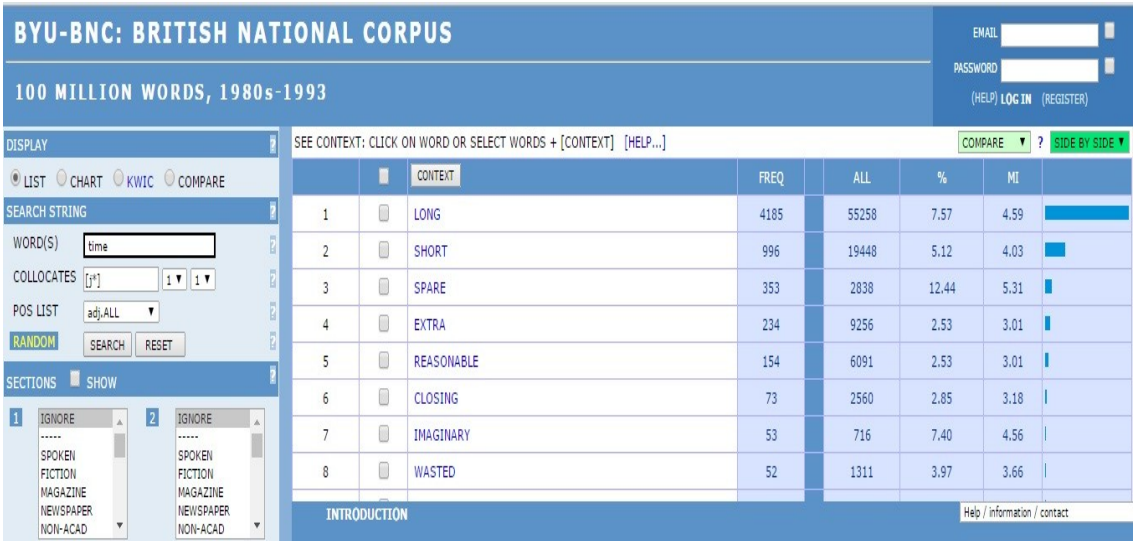

Рис. 3. Данные корпуса о сочетаемости лексической единицы time по модели adj + time

2) в сочетании с прилагательными (adj + time) данная лексическая единица имеет высокоустойчивые сочетания: **long time, short time, spare time, extra time, reasonable time, extra time**;

3) в сочетании с существительными (time + noun) данная лексическая единица имеет высокоустойчивые сочетания: **time limit, time span, time lap, time warp**.

Подобный анализ сочетаемости лексических единиц можно провести для всех частей речи, как главных, так и служебных. Приведенные примеры показывают эффективную возможность использования лингвистических корпусов и статистического метода для выделения устойчивых сочетаний.

Список литературы

1. *Дерябина И.В.* Обучение управлению английских глаголов на основе английского национального корпуса как педагогическая проблема // Вестник Тамбовского университета. Серия Гуманитарные науки. Тамбов, 2012. Вып. 10 (114). С. 154-158.
2. *Рязанова Е.А.* Формирование грамматических навыков речи в целях обучения иностранно-

му языку в языковом вузе // Вестник Тамбовского университета. Серия Гуманитарные науки. Тамбов, 2012. Вып. 9 (113). С. 153-156.
3. *Сысоев П.В.* Лингвистический корпус, корпусная лингвистика и методика обучения иностранным языкам // Иностранные языки в школе. 2010. № 5. С. 12-21.
4. *Сысоев П.В., Кокорева А.А.* Обучение студентов профессиональной лексике на основе корпуса параллельных текстов // Язык и культура. 2013. № 1 (21). С. 114-124.
5. Курс лекций по корпусной лингвистике. URL: http://rykov-cl.narod.ru (дата обращения: 01.02.2016).
6. *Иорданская Л.Н., Мельчук И.А.* Смысл и сочетаемость в словаре. М., 2007.
7. *Ягунова Е.В., Пивоварова Л.М.* Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов // Научно-техническая информация. Серия 2. Информационные процессы и системы. М., 2010. № 6.
8. *Ахманова О.С.* Словарь лингвистических терминов. М., 1966.
9. *Борисова Е.Г.* Что такое коллокации и как их изучать. М., 1995.
10. *Баранов А.Н., Добровольский Д.О.* Аспекты теории фразеологии. М., 2008.
11. *Влавацкая М.В.* Теоретические основы комбинаторной лингвистики: лексикологический и лексикографический аспекты: автореф. дис. … д-ра филол. наук. Новосибирск, 2013.
12. *Мельчук И.А.* О терминах «устойчивость» и «идиоматичность» // Вопросы языкознания. 1960. № 4.
13. *Захаров В.П.* Корпусная лингвистика. СПб., 2005.

References

1. Deryabina I.V. Obuchenie upravleniyu angliyskikh glagolov na osnove angliyskogo natsional'nogo korpusa kak pedagogicheskaya problema. *Vestnik Tambovskogo universiteta. Seriya Estestvennye i tekhnicheskie nauki – Tambov University Reports. Series: Natural and Technical Sciences.* Tambov, 2012, no. 10 (114), pp. 154-158.
2. Ryazanova E.A. Formirovanie grammaticheskikh navykov rechi v tselyakh obucheniya inostrannomu yazyku v yazykovom vuze. *Vestnik Tambovskogo universiteta. Seriya Estestvennye i tekhnicheskie nauki – Tambov University Reports. Series: Natural and Technical Sciences.* Tambov, 2012, no. 9 (113), pp. 153-156.
3. Sysoev P.V. Lingvisticheskiy korpus, korpusnaya lingvistika i metodika obucheniya inostrannym yazykam. *Inostrannye yazyki v shkole – Foreign languages for schools*, 2010, no. 5, pp. 12-21.
4. Sysoev P.V., Kokoreva A.A. Obuchenie studentov professional'noy leksike na osnove korpusa parallel'nykh tekstov. *Yazyk i kul'tura – Language and Culture*, 2013, no. 1 (21), pp. 114-124.
5. *Kurs lektsiy po korpusnoy lingvistike*. Available at: http://rykov-cl.narod.ru (accessed 01.02.2016).
6. Iordanskaya L.N., Mel'chuk I.A. *Smysl i sochetaemost' v slovare*. M., 2007.
7. Yagunova E.V., Pivovarova L.M. Priroda kollokatsiy v russkom yazyke. Opyt avtomaticheskogo izvlecheniya i klassifikatsii na mate-riale novostnykh tekstov. *Nauchno-tekhnicheskaya informatsiya. Seriya 2. Informatsionnye protsessy i sistemy – Automatic Documentation and Mathematical Linguistics*, Moscow, 2010, no. 6.
8. Akhmanova O.S. *Slovar' lingvisticheskikh terminov*. Moscow, 1966.
9. Borisova E.G. *Chto takoe kollokatsii i kak ikh izuchat'*. Moscow, 1995.
10. Baranov A.N., Dobrovol'skiy O.G. *Aspekty teorii frazeologii*. Moscow, 2008.
11. Vlavatskaya M.V. *Teoreticheskie osnovy kombinatornoy lingvistiki: leksikologicheskiy i leksikograficheskiy aspekty*. Avtoreferat dissertatsii … doktora filologicheskikh nauk. Novosibirsk, 2013.
12. Mel'chuk I.A. O terminakh “ustoychivost'” i “idiomatichnost'”. *Voprosy yazykoznaniya – Voprosy Jazykoznanija (Topics in the study of language)*, 1960, no. 4.
13. Zakharov V.P. *Korpusnaya lingvistika*. St. Petersburg, 2005.

Поступила в редакцию 18.03.2016 г.
Received 18 March 2016

UDC 159.946.3

COLLOCATIONS STUDY BASED ON THE TEXT CORPORA

Tatyana Yurevna PAVELEVA, Lipetsk State Pedagogical University, Lipetsk, Russian Federation, Candidate of Pedagogy, Associate Professor, Linguistics and Intercultural Communication Department, e-mail: tatianapavelieva@live.com

The problem of extracting and studying collocations on the basis of linguistic corpora is considered. Such concepts of corpus linguistics as representativeness and the corpus volume, corpus manager (Concordance) and text corpora peculiarities are studied. The examples of definitions of collocations are given from the point of view of traditional and corpus linguistics. The definition of collocation as statistically stable combinations of lexical units is especially underlined. Different approaches to the definition of context as a base of studying collocation are examined. Such advantages of using text corpora in the linguistic studies as comprehensiveness and relevance of textual information are studied as well as the special statistic methods of processing information. This method is called MI (mutual information) and it compares dependent context-related frequencies with independent ones and allows single out all possible collocations. The examples of extracting collocation of the word time are set.

Key words: collocation; corpus linguistics; text corpora; context; MI (mutual information).

DOI: 10.20310/1810-0201-2016-21-3/4(155/156)-56-61