

Павловская Ирина Юрьевна

Доктор филологических наук,
профессор, Санкт-Петербургский
государственный университет
(Санкт-Петербург, Россия)
E-mail: pavlovskayairina2@gmail.com

Pavlovskaya Irina Yurievna

Doctor of Philology, professor,
St. Petersburg State University
(Saint-Petersburg, Russia)
E-mail: pavlovskayairina2@gmail.com

Горина Ольга Григорьевна

Кандидат педагогических наук,
Санкт-Петербургской школы социальных
и гуманитарных наук Национального
исследовательского университета «Высшая
школа экономики»
(Санкт-Петербург, Россия)
E-mail: gorina@bk.ru

Gorina Olga Grigorievna

Ph.D. (Education), Senior Lecturer, National
Research University Higher School of
Economics, of the Department of Foreign
Languages, St. Petersburg School of Social
Sciences and Humanities
(Saint-Petersburg, Russia)
E-mail: gorina@bk.ru

**КОРПУСНО-КОГНИТИВНЫЕ МЕТОДЫ
ИЗУЧЕНИЯ ЛЕКСИЧЕСКОЙ СТОРОНЫ
РЕЧИ НА ИНОСТРАННОМ ЯЗЫКЕ**

**CORPUS-BASED COGNITIVE ORIENTED
METHODS OF LEXICAL ANALYSIS IN
FOREIGN LANGUAGE STUDIES**

Аннотация. В статье рассматриваются способы анализа текстов и лексических единиц, выделяемых на основе статистического аппарата корпусного менеджера, таких как коллокации, коллигации, чанки, кластеры, элементы хеджинга и др. Представлена когнитивная интерпретация этих единиц. Выводы основаны на результатах анализа авторского корпуса в 1 млн 700 тыс. словоупотреблений. Излагаются способы использования и характеристики основных корпусных инструментов анализа: конкордансов, коллокационно-коллигационных профилей, частотных характеристик, поиска ключевых слов, облаков слов и др. В качестве справочного источника используется Британский национальный корпус. Технически исследование выполнено с использованием корпусного менеджера WordSmith Tools 6.0.

Summary. The paper is concerned with the methods of analysis of texts and lexical units such as collocations, colligations, chunks, clusters, hedging, etc. The aim of the research is to equip linguists, teachers and students with the skills of objective language analysis and adequate use of modern English language. The research is based upon statistical instruments of corpus manager. Cognitive interpretation of the data is presented. Conclusions are based on the results of compiling an original corpus for teaching British Studies, including 1 700 000 entries, and on analyzing professionally significant ones in a data-driven way. Implementation of corpus instruments such as concordances, collocation and colligation profiles, statistics, frequency characteristics, key-words search, word-clouds and others are under consideration. British National Corpus is used as a reference tool. Technically the research is fulfilled with the help of corpus manager WordSmith Tools 6.0.

Ключевые слова. Корпус, корпус-менеджер, лингвостатистика, коллокации, коллигации, чанки, хеджинг.

Key words. Corpus, corpus manager, linguostatistics, collocations, colligations, chunks, hedging.

Введение

В современном языкознании начинает складываться понимание роли корпусных технологий в когнитивном, лингвистическом и педагогическом контекстах. Предмет настоящего исследования обусловлен необходимостью выявления роли корпусных

баз данных в исследовательских задачах этих направлений и их прикладных возможностях. Уже существуют практически реализованные корпусные проекты, выведенные в современное информационное пространство [7], [6], [3].

Языковедам, преподавателям, лингводидактам необходим обмен знаниями по практическим вопросам о преимуществах и недостатках корпусной лингвостатистики, характеристикам лингвостатистических показателей, реализуемых в корпус-менеджерах и интерфейсах.

Основная часть

Исследования в области прикладной лингвистики, и особенно в области корпусной лексикографии, привели к тому, что наметились лексически-ориентированные подходы к описанию характера речепорождения: «...многие синтаксические явления стали рассматриваться как проекция общих закономерностей функционирования лексикона» [1, с. 263]. Дж. Синклер в своих многочисленных работах также придерживается той точки зрения, что лексика, а не синтаксис, отвечает за организацию и структуру (patterning) речи [16], [17].

Вместе с тем, такие свойства современных корпусных технологий, как лингвистическая и статистическая наглядность, широкие поисковые возможности, репрезентация различных регистровых особенностей, вполне применимы при отборе, например, профессиональных лексических минимумов и репрезентативного лингвистического материала при создании корпусных грамматик и пособий, учебных материалов, тестов. В контексте методического применения корпусные ресурсы также дают возможность преподавателю работать в самых современных форматах, количественно и когнитивно согласованных с задачами обучения.

Следует отметить, что идея использования корпусных данных в лингводидактическом исследовании зиждется на убеждении, что внимание исследователя должно быть сосредоточено на закреплённых в речевом сообществе способах номинации, т.е. на реально функционирующем языке. Сказанное означает, что в любом корпусном исследовании значительно место когнитивной составляющей. Обращаясь к корпусной лингвостатистике, коллокационно-коллигационному профилю слов, мы, прежде всего, можем количественно оценить определённую когнитивную реальность.

Так, например, по данным грамматики издательства «Пирсон», созданной на основе анализа 40-миллионного корпуса устной и письменной английской речи, слова, встречающиеся один раз на 1.000 словоупотреблений (т.е. высокочастотные слова), в живой речи можно слышать один раз в 8,5 минут, а слова, встречающиеся один раз на 100.000 словоупотреблений (также достаточно высокочастотные слова) – один раз в 14 часов [5]. Когнитивная реальность складывается из следующей статистики. Один миллион словоупотреблений в устных речевых актах составляет 120 часов. Взрослый говорящий принимает участие в речевых актах, состоящих из 15 млн слов в год, а человек старше 50 лет в среднем слышал или принимал участие в актах речи, состоящих из более, чем полумиллиарда словоупотреблений [19].

То обстоятельство, что статистически язык можно представить большим набором редких событий, имеет далеко идущие последствия для понимания когнитивных процессов, происходящих при порождении речи. Под влиянием корпусных исследований переосмысливаются количественные и качественные показатели словарного запаса говорящего, необходимого для успешной речевой деятельности. При этом корпусный отбор может обеспечить надёжные и достоверные статистические опоры, позволяет достичь большей точности и сосредоточенности на ключевых понятиях и

лексических единицах избранной предметной области, которые несут основную смысловую нагрузку и обладают реальной практической значимостью.

Для обучения студентов специальности «Британское регионоведение» нами был собран корпус текстов в 1 млн 700 тыс. словоупотреблений и проведен отбор тематически-значимой лексики (ключевых слов) на основе статистической процедуры отбора. Обработка текстов осуществлялась с помощью корпус-менеджера «Уорд Смит Тулз», версия 6.0 (WordSmith Tools 6.0) [14]. Данный инструментарий хорошо подходит для выполнения таких задач, как составление списков частотности, конкордансов и поиска ключевых слов*. Следует отметить, что сегодня поиск ключевых слов как статистической опоры при отборе наиболее релевантного кодабулара становится стандартной процедурой обработки корпусной базы, основанной на сравнении двух корпусов текстов: опорного (или справочного) и изучаемого. Кроме того, корпус-менеджер согласован с разметкой лингвистической/текстовой базы Британского национального корпуса (БНК), что позволило нам использовать и его 100-миллионную текстовую базу в качестве справочного корпуса. Ключевые слова, полученные с помощью автоматического корпусного отбора, могут далее изучаться, например, на предмет их коллокационно-коллигационного рисунка, наиболее вероятного грамматического оформления и т.д. с учетом различных корпусных параметров в зависимости от задач исследователя. В частности, корпусный отбор ключевых слов может найти свое применение в составлении первичного словника при формировании поуровневого словаря для изучающих профессионально-ориентированный иностранный язык.

Вместе с тем, сведения о лексических единицах, а не только о словах (а именно, коллокациях, коллигациях, чанках, речевых стереотипах, кластерах, явлении хеджинга и др.), и идеи, которые развиваются в рамках когнитивных течений в отношении речепорождения, также являются, с нашей точки зрения, предметом корпусного исследования. Корпусный инструментарий позволяет выделять такие построения на основе частотности на больших массивах языковых данных и интерпретировать результаты [5], [13] и др.

Наличие в памяти говорящего большого количества высоковероятных готовых словосочетаний и частей предложений увеличивает возможность более легких оперативных действий при их комбинировании. Эти блоки, или чанки, языка аранжируются в соответствии с речедетельностной задачей говорящего и одновременно с его речемыслительной деятельностью.

Термин *chunk / chunking* («чанк» / «образование чанков») использовал американский психолог Д. Миллер для характеристики кратковременной памяти человека [12]. Понятие «чанк» также широко употребляется в теориях обучения. В современных классификациях лексических единиц в русле корпусных исследований «чанк» рассматривается как обобщенный, зонтичный термин, который включает в себя все другие термины, которыми обозначаются такие сочетания слов как коллокации, лексические пучки (*lexical bundles*) [5] и т.п.

В соответствии с когнитивными теориями речепорождения представляется маловероятным, чтобы выражение, например, *It takes a bit of getting used to*, постоянно конструировалось говорящим по грамматическим правилам. Вероятнее всего, фрагмент из восьми слов хранится и извлекается из памяти целиком, а при необходимости подвергается лишь незначительной грамматической коррекции.

* При выявлении КС выполняется расчет по критерию χ -квадрат (хи-квадрат, *chi-square*) или тест на логарифмическое правдоподобие (*log-likelihood test*) (Scott 2012)

Обратимся к некоторым деталям корпусного поиска таких построений и выявления их коллокационно-коллигационного потенциала/рисунка. К корпусным инструментам традиционно относится генерирование строк конкорданса, или строк сочетаемости, изучение которых позволяет наблюдать за лексико-грамматическим профилем лексической единицы, найти левые и правые коллокаты визуально, т.е. вертикально сканируя конкорданс в формате «картинки». Так, например, для исследования комбинаторного потенциала частотных сегодня новостных выражений *on the table / off the table* (выносить вопрос на обсуждение и снимать вопрос с повестки дня) достаточно двух десятков строк сочетаемости, сгенерированных корпус-менеджером. Становится очевидным, что их коллокационный потенциал развивается на основе метафоры в сочетании с такими глаголами как *put* (*on the table*), *be* (*on the table*), *lay* (*on the table*), *take off* (*the table*), *sweep off* (*the table*) – положить, убрать, смести (со стола). Изучение конкордансов способствует сознательному торможению влияния родного языка на основе возможной интерференции, помогает избежать нерегулируемых аналогий и, если нет упроченного в прошлом стереотипа сочетания слов, действует как профилактическое предупреждение в отношении сочетаемости слов. Вместе с тем исследователь может сортировать слова в зависимости от нужных ему контекстов, изучать частотность слова по жанрам корпуса.

Автоматический корпусный поиск коллокаций основан на понимании того факта, что коллокация определяется статистически как более высокая совместная встречаемость слов, отличная от случайной. Корпусные инструменты позволяют не только найти коллокат, но и специфицировать часть речи при поиске коллокаций. Как правило, одним из вариантов сортировки является показатель взаимной информации (Mutual Information, MI). Это автоматически рассчитываемый в корпус-менеджерах параметр, указывающий на возможную сочетаемость. Так, словосочетание *solve problem* имеет высокий показатель MI, а *solve the* – низкий. При расчете показателя взаимной информации учитываются следующие статистические факты. Артикль *the* встречается в окрестности (2–3 слова) глагола *solve* значительно чаще в силу высокой абсолютной частотности определенного артикля, который выходит на первое место по частотности в подавляющем большинстве текстов. Этот метод, однако, тоже имеет свой недостаток. При расчете описываемого параметра приоритет могут получать необычные слова, что приводит к тому, что необычные комбинации слов попадают в выдачу с высоким показателем MI, притом, что такие комбинации могут быть специфичны только для определенного корпуса [10].

При обнаружении чанков важной оказывается возможность обозначить зону поиска, т.е. в какой именно удаленности от слова мы ищем коллокат. Как правило, можно задать зону поиска: от 1 до 5 слов, слева (L) или справа (R) от изучаемого слова. Интересно отметить, что при поиске ближайших коллокатов, например, из числа прилагательных к существительному *day* мы получим частотное словосочетание *summer day*. Если же расширить поиск до окрестности в 2–4 слова от изучаемого, то есть вероятность обнаружить чанк, такой как *beautiful summer day*. Сочетание корпусного поиска коллокаций и конкорданса дает возможность найти более длинные, но вместе с тем, рекуррентные чанки, как например: *(it) takes a bit of getting used to, there is a certain stigma attached to*, или выражения со словами *thanks* и *efforts* (которые по показателю MI частотны вместе): *thanks to the efforts of the, thanks for your efforts in, thanks to her own efforts, thanks largely to the efforts of, thanks for the efforts they make* — в общей сложности 46 примеров по БНК, 20 из которых только на сочетание *thanks to the efforts*. Возможность поиска статистически значимых комбинаций слов особенно важна для не-носителей английского языка.

Многочисленные контексты дают возможность наблюдать за статистическими проявлениями явления *семантической просодии*, под которой понимается тяготеение к определенным типам контекстов (положительно или отрицательно коннотированным словам) [2]. Например, отмечается, что в более чем 90 % случаев слово *cause* встречается в негативных контекстах, с негативно коннотированными словами: *cancer, commotion, crisis, delay*. Напротив, слово *provide* приводится как типичное для положительных контекстов, в наиболее типичных сочетаниях со словами: *care, food, help, jobs, relief, support* [18]. Кластер *something of a* в 50 % случаев выступает в препозиции к отрицательно коннотированным существительным (*something of a stigma, disappointment, two-edged sword*). До компьютеризированного языкового анализа этот феномен не был систематизирован с подтверждением на аутентичных материалах.

Исследование этого феномена также тесно связано с изучением стратегии хеджинга – социо-лингвистического явления, которое проявляет себя на разных языковых уровнях и чрезвычайно многообразно [4], [11], [8]. Корпус позволяет выработать методологию исследования данного явления и упрощает поиск доказательной базы. Например, лексический пучок *something of a* встречается в БНК 1,155 раз. Возникает убежденность, что такая высокая частотность кластера говорит о важных прагматических функциях, закрепленных за этой цепочкой слов. Корпусный поиск помогает выявить, изучить и систематизировать сведения о лексических пучках. Изучение контекстов употребления упомянутого кластера показывает, что лексический пучок *something of a* выступает в роли средства хеджинга, смягчает резкость существительных с отрицательной коннотацией, а также чаще встречается в письменной речи: *something of a disappointment* (*некоторое разочарование*); *something of a stigma in being an unemployed professional* (*весьма постыдно иметь профессию и быть безработным*); *something of a predicament* (*несколько неприятная ситуация*); *something of a return to “normal”* (*некое подобие* возвращения к «норме»/нормальной ситуации). Самыми частотными правыми коллокатами *something of a* являются существительные: *shock, surprise, mystery, problem, misnomer, disappointment, challenge, dilemma, reputation, puzzle*. Кластер, или лексический пучок, представлен со сравнимой частотностью в академическом, новостном, неакадемическом письменных регистрах, но не частотен в устном разговорном регистре БНК [9].

В поле зрения когнитивных лингвистов сегодня все чаще и чаще попадают лингвистические явления, характерные только для устных или только для письменных жанров как еще одно отражение когнитивной речевой реальности. В корпусных исследованиях дихотомия «устная / письменная речь» рассматривается с точки зрения лексической плотности, т. е. удельного количества содержательных слов на предложение, наличия включений (*erm, right, yeah, please*), дейктических средств, маркеров интерактивности (*I, You, yeah*), сочетаний существительных, союза *that*. При этом самый высокий процент содержательных слов приходится на новостные тексты, а самый низкий – на записи спонтанной устной (обиходно-бытовой) диалогической речи [5]. Очевидно, что меньшая лексическая плотность устной речи является в первую очередь отражением того факта, что говорящий не может «онлайн», или в «живом эфире» производить такой же лексически плотный поток речи, какой характерен для письменных жанров. Кроме того, определенную роль играет включенность ситуации в контекст общения, что приводит к повышенному содержанию дейктических средств [13].

Корпусные исследования на примере союза *that* в контрастирующих контекстах устной и письменной речи говорят о том, что самые частотные употребления *that*, судя по разметке корпуса БНК, в письменной речи приходятся на случаи функцио-

нирования в качестве союза, определителя, наречия, т.е. для дальнейшего развития аргументов в сложных придаточных предложениях, в то время как в устной речи *that* отсылает к разделенному знанию (shared knowledge) [15].

Выводы

Таким образом, сегодня использование корпуса и развитие корпусной компетенции исследователя (умения использовать корпусную справочную систему) можно считать уже обязательным форматом представления языковых данных, наряду с традиционными грамматиками и словарями. Он является залогом достоверности и адекватности лингвистического исследования.

Литература

1. Гвишиани Н.Б. Введение в контрастивную лексикологию (Англо-русские межъязыковые соответствия). М.: Либроком, 2010. 288 с.
2. Гвишиани Н.Б. Практикум по корпусной лингвистике. М.: Высшая школа, 2008. 191 с.
3. Камшилова О.Н., Колина М.В., Николаева В. А., Осипова А. В., Разумова В. В. SPb EFL Learner Corpus (2008). <http://www.spbeflcorp.ru/about.html> (дата обращения 01.10. 2014)
4. Марюхин А.П. Непрямая коммуникация в научном дискурсе (на материале русского, английского, немецкого языков): автореф. дис.... канд. филол. наук. М., 2010. 24 с.
5. Biber D., Johansson S., Leech G., Conrad S., Finegan E. Longman Grammar of Spoken and Written English. – S.I.: Pearson Education Limited, 1999. 1204 p.
6. Braun S., Kohn K., Lenz C., Sieber G. Widmann H. ELISA (English Language Interview Corpus as a Second Language Application). Tuebingen, 2003–2004. URL: http://www.uni-tuebingen.de/elisa/html/elisa_index.html (дата обращения 1.12.2014)
7. Chukharev-Hudilainen E., Klepikova T. A. Linguatorium in Teaching English Writing for Specific Purposes / Technology for Second Language Learning. Iowa State University, 2012.
8. Crystal D. On keeping one's hedges in order // English Today. 1988. № 15. Pp. 46–7.
9. Davies M. (2004) BYU-BNC. (Based on the British National Corpus from Oxford University Press). URL: <http://corpus.byu.edu/bnc/>.
10. Hunston S., Laviosa S. Corpus Linguistics. – Birmingham: Centre for English Language Studies. The University of Birmingham, 2000. 112 p.
11. Lakoff D. Hedges: A study in meaning criteria and the logic of fuzzy concepts // Journal of Philosophical Logic. 1972. № 2(4). Pp. 458–508. DOI: 10.1007/BF00262952.
12. Miller G.A. The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information // Psychological Review. 1956. № 63. Pp. 81–97. URL: <http://psychclassics.yorku.ca/Miller/>.
13. O’Keeffe A., McCarthy M., Carter R. From Corpus to Classroom: language use and language teaching. Cambridge: Cambridge Univ. Press, 2007. 315 p.
14. Scott M. Wordsmith Tools: Software. Oxford: Oxford University Press, 2012.
15. Scott M., Tribble C. Textual Patterns: key words and corpus analysis in language education: Studies in Corpus Linguistics. Amsterdam/Philadelphia: John Benjamins, 2006. 200 p.
16. Sinclair J. Corpus: Concordance and Collocation. Oxford: Oxford Univ. Press, 1991. 137 p.
17. Sinclair J. The search for units of meaning // Textus. 1996. № 9 (1). Pp. 75–106.
18. Stubbs M. Corpus evidence for norms of lexical collocation // Cook G., Seidlhofer B. (eds.) Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson. — Oxford: Oxford Univ. Press, 1995. P. 245–56.
19. Teaching Collocation. Further developments in Lexical Approach / Editor Michael Lewis. Thompson Heinle: 2000.

References

1. Gvishiani N.B. *Vvedenie v kontrastivnuu leksikologiu (Anglo-russkie mezhyazykovye sootvetstviia)* [Introduction to Contrastive Lexicology: English-Russian cross-language correspondences]. Moscow, 2010. 288 p.

2. Gvishiani N.B. *Praktikum po korpusnoi lingvistike: Ucheb. posobie po angliiskomu iazyku* [Manual in corpus linguistics: Teaching material for the English language study]. Moscow, 2008. 191 p.
3. Kamshilova O.N., Kolina M.V., Nikolaeva V.A., Osipova A.V., Razumova V.V. *SPb EFL Learner Corpus* (2008). Available at: <http://www.spbeflcorp.ru/about.html> (assessed 01.10. 2016)
4. Mariuhin A.P. *Nepriamaia kommunikaciia v nauchnom diskurse (na materiale russkogo, angliiskogo, nemeckogo iazykov) : Avtořef. dis.... kand. filol. nauk* [Mediated communication in scientific discours, Ph.D. dissertation (synopsis)]. Moscow, 2010. 24 p.
5. Biber D., Johansson S., Leech G., Conrad S., Finegan E. *Longman Grammar of Spoken and Written English*. – S.l.: Pearson Education Limited, 1999. 1204 p.
6. Braun S., Kohn K., Lenz C., Sieber G. Widmann H. *ELISA (English Language Interview Corpus as a Second Language Application)*. Tuebingen, 2003–2004. Available at: http://www.uni-tuebingen.de/elisa/html/elisa_index.html (дата обращения 1.12.2014)
7. Chukharev-Hudilainen E., Klepikova T. A. *Linguatorium in Teaching English Writing for Specific Purposes / Technology for Second Language Learning*. Iowa State University, 2012.
8. Crystal D. On keeping one's hedges in order. *English Today*, 1988, no. 15, pp. 46–7.
9. Davies M. (2004–) *BYU-BNC. (Based on the British National Corpus from Oxford University Press)*. Available at: <http://corpus.byu.edu/bnc/>.
10. Hunston S., Laviosa S. *Corpus Linguistics*. Birmingham: Centre for English Language Studies. The University of Birmingham, 2000. 112 p.
11. Lakoff D. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 1972, no. 2(4), pp. 458–508. DOI: 10.1007/BF00262952.
12. Miller G.A. The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review*, 1956, no. 63, pp. 81–97. Available at: <http://psych-classics.yorku.ca/Miller/>.
13. O'Keeffe A., McCarthy M., Carter R. *From Corpus to Classroom: language use and language teaching*. Cambridge: Cambridge Univ. Press, 2007. 315 p.
14. Scott M. *Wordsmith Tools: Software*. Oxford: Oxford University Press, 2012.
15. Scott M., Tribble C. *Textual Patterns: key words and corpus analysis in language education: Studies in Corpus Linguistics*. Amsterdam ; Philadelphia: John Benjamins, 2006. 200 p.
16. Sinclair J. *Corpus: Concordance and Collocation*. Oxford: Oxford Univ. Press, 1991. 137 p.
17. Sinclair J. The search for units of meaning. *Textus*, 1996, no. 9 (1), pp. 75–106.
18. Stubbs M. Corpus evidence for norms of lexical collocation. Cook G., Seidlhofer B. (eds.) *Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson*. Oxford: Oxford Univ. Press, 1995, pp. 245–56.
19. Teaching Collocation. *Further developments in Lexical Approach ; Ed. Michael Lewis*. Thompson Heinle: 2000.

Павловская И.Ю., Горина О.Г. Корпусно-когнитивные методы изучения лексической стороны речи на иностранном языке // Вестник Череповецкого государственного университета. 2017. №1. С. 132–138.

For citation: Pavlovskaya I.Yu., Gorina O.G. Corpus-based cognitive oriented methods of lexical analysis in foreign language studies. *Bulletin of the Cherepovets State University*, 2017, no. 1, pp. 74–80.