УДК 81'374

О НЕКОТОРЫХ ВОЗМОЖНОСТЯХ ВЫЯВЛЕНИЯ КОЛЛОКАЦИЙ С ПОМОЩЬЮ ИНТЕРНЕТ-ТЕХНОЛОГИЙ

Д. В. Андрианова

Институт лингвистических исследований РАН

Поступила в редакцию 8 апреля 2019 г.

Аннотация: *современные научные исследования, в том числе филологические, требуют применения современных методов работы, что связано с необходимостью эффективного использования Интернет-источников. В статье анализируется ряд ресурсов, которые могут быть полезными в лексикографической работе для поиска и отбора наиболее релевантных коллокаций из массива Интернет-текстов.*
Ключевые слова*: лексикография, коллокации, Интернет, корпуса русского языка.*

Abstract*: modern scientific research, also in philology, requires the use of modern methods of work, which involve effective use of the Internet sources. The article analyzes a number of sources that may be useful in lexicographical work for searching and selecting the most relevant contexts for the collocations from Internet texts.*
Keywords*: lexicography, collocations, Internet, Russian language corpuses.*

Благодаря развитию Интернета и Интернет-технологий карточные каталоги в работе лексикографа уступают место работе с электронными текстами. Именно поэтому особенно актуальным в лексикографической практике становится умение эффективно организовать автоматический поиск и отбор необходимых и релевантных единиц. Одной из интересных и актуальных задач в контексте оптимизации поиска в Интернет является выявление т.н. коллокаций, т.е. сочетаний слов, которые характеризуются относительно устойчивой частотностью совместной встречаемости. Такие сочетания слов присущи всем естественным языкам [1 с. 59], однако их изучение было начато только во второй половине XX в. в работах J. R. *Firth, И. А. Мельчука,* О. С. Ахмановой, С. Г. Тер-Минасовой и др. Для словарной работы коллокации интересны тем, что «они являются периферийными единицами как для традиционной лексикологии, которая описывает в основном свободную сочетаемость лексем, так и для фразеологии, занимающейся непосредственно идиомами. Исходя из этого, коллокации, оказываясь на границе между лексикологией и фразеологией, занимают промежуточное положение в системе языка» [1 с. 57]. Одной из задач современной практической лексикографии, по всей видимости, является поиск и фиксация коллокаций для целей последующего изучения этого языкового феномена.

Сегодня «все лингвисты, работающие в самых разных направлениях, как правило, проводят свои исследования на базе корпусов» [2, с. 20]. Вероятно, самый известный и часто привлекаемый для исследовательских целей корпус в России — **НКРЯ** (http://www.ruscorpora.ru/). В отличие от более крупных корпусов, этот корпус создается «лингвистами по заранее описанной технологии, которую можно назвать «классической». Для подобных корпусов тексты отбираются, размечаются и далее загружаются в корпус» [3 с. 74]. Одним из больших преимуществ этого корпуса, по сравнению с автоматически собранными корпусами, является то, что все тексты в нем созданы русскоязычными авторами на русском языке. Удобный интерфейс позволяет легко осуществлять простейшую выборку по времени написания текста, по жанру (художественная литература, публицистика и т.д.). Конструктор запросов также позволяет уточнять грамматические формы слов и осуществлять поиск коллокаций с дистантным расположением компонентов. На стадии разработки находится синтаксический корпус, с помощью которого можно очень детализированно задать синтаксическую структуру коллокации. Однако в настоящий момент объем синтаксического корпуса составляет всего 100 лексических функций, которым соответствует около 21 тысячи словосочетаний.

Более объемные, по сравнению с НКРЯ, корпуса составляются в большинстве случаев автоматически из текстов, полученных из Интернет, которые затем обрабатываются специальными программами с целью удаления дублей и повторов, выполнения морфологической и морфосинтаксической разметки и проч. [3, с. 74]. К этому типу относится корпус русскоязычных текстов из художественной литературы **Google N-gram Viewer** (https://books.google.

com/ngrams). Этот ресурс с лаконичным интерфейсом позволяет моментально отобразить на графике количество употреблений какого-либо слова или сочетания слов за период с 1800 по 2009 гг. Что особенно удобно, под графиком можно выбрать интересующий пользователя период и просмотреть источники с цитатами, в которых встретилось искомое сочетание. Чтобы отобразить наиболее часто встречающиеся рядом с определенным словом (словоформой) слова, необходимо поставить знак звездочки до или после этого слова. Однако в данном случае будут отображены только слова, стоящие непосредственно перед этой единицей или после нее. И хотя среди этих оборов могут быть найдены коллокации, специального инструмента для поиска коллокаций, в т.ч. с дистантным расположением компонентов, в Google N-gram Viewer нет.

Еще один большой созданный автоматически корпус русскоязычных текстов — **RuTenTen** (www.sketchengine.eu/rutenten-russian-corpus/). Преимуществом этого корпуса является наглядное отображение частотных для заданного слова синтаксических моделей и коллокаций, распределенных по этим моделям. При этом нельзя задать поиск конкретной коллокации. К недостаткам этого корпуса можно отнести отсутствие хронологической сортировки в отображении конкорданса, а также включение в него как русскоязычной, так и переводной литературы (даже при заданном поиске по сайтам с доменом.ru отсортировать переводные тексты не представляется возможным). Отметим также сложность атрибуции: в конкордансе дается только ссылка на сайт — источник цитаты, но не указывается автор и название текста.

Много интересных возможностей открывает **ГИКРЯ (**http://www.webcorpora.ru/), материалом для которого являются тексты крупнейших ресурсов social media, а также новостных материалов и контента «Журнального зала». Огромным преимуществом этого корпуса является то, что тексты в нем принадлежат современным носителям языка и отражают актуальные языковые процессы. Все сниппеты, т.е. контексты, отобранные по запросу, четко атрибутируются по году написания текста, году рождения автора и т.д. Благодаря очень детализированной разметке и конструктору запросов пользователь может задавать очень точные запросы для поиска коллокаций, построенных по определенной синтаксической модели.

Еще одна возможность выявления коллокаций с помощью Интернет-ресурсов — это поиск по пользовательским запросам (логам) [4 с. 223]. Сервис **Google Trends** (http://www.google.com/trends/) в результатах по запросу какого-либо слова в графах «еще по теме» и «похожие запросы» выдает наиболее частотные мини-контексты, в которых это слово встречается в поисковых запросах пользователей Интернет. Для решения той же самой задачи в **Яндекс.Статистика** (http://wordstat.yandex.ru/) необходимо ввести слово, для которого подбираются коллокации, в кавычках дважды через пробел для получения двухсловных запросов и трижды через пробел для получения трехсловных.

Помимо поиска в корпусах и по логам можно отметить ресурс **Яндекс.Блоги** (https://yandex.ru/blogs/). Основным преимуществом выборки по блогам является, безусловно, актуальность материала. Именно этот инструмент помогает отследить узуальные особенности словоупотребления и значений конкретного слова или коллокации, поскольку блогосфера в отличие, например, от большей части корпусов, отражает письменную речь носителей языка вне сферы профессионального «писательства». В числе безусловных преимуществ поиска по блогосфере является отсутствие переводных текстов. Применив настройки расширенного поиска, можно выбрать регион и период поиска.

Подводя итог представленному выше обзору некоторых возможностей поиска коллокаций в текстах Интернет, нужно отметить, что нельзя назвать один ресурс, пользование которым позволит решить любую задачу в заданных рамках. Так, для наиболее общего представления о поведении искомой единицы в контексте лучше всего подойдет НКРЯ, современные значения также будут отражены в ГИКРЯ. Поиск по логам и блогам дает возможность проследить современные тенденции в семантике и употреблении коллокаций. Поиск характерных синтаксических конструкций с заданным словом оптимально осуществлять в системе RuTenTen11. Употребительность той или иной единицы можно проследить с помощью сервисов Яндекс.Статистика и Google Trends. Каждый из рассмотренных ресурсов имеет свою специфику, знание которой облегчит пользователю задачу максимально быстро и эффективно найти коллокации и составить представление об особенностях их семантики и стилистики в соответствующих контекстах.

ЛИТЕРАТУРА

1. Влавацкая М. В. Комбинаторная лексикология: функционально-семантическая классификация коллокаций / М. В. Влавацкая // Филологические науки. Вопросы теории и практики.— 2015.— № 11/1 (53).— С. 56–60.

2. Захаров В. П. Корпуса русского языка / В. П. Захаров // Труды института русского языка им. В. В. Виноградова.— Т. 6.— 2015.— С. 20–65.

3. Хохлова М. В. Обзор больших русскоязычных корпусов текстов. / М. В. Хохлова // Компьютерная лингвистика и вычислительные онтологии: сборник научных статей. Труды XIX Международной объединенной научной конференции «Интернет и современное общество» (IMS-2016), СПб., 22–24 июня 2016 г.— 2016.— С. 74–77.

4. Словарь бытовой терминологии: новые проблемы и новые методы / Б. Л. Иомдин [и др.] // Компьютерная

лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая-3 июня 2012 г.).— Вып. 11.— 2012.— С. 213–226.

Институт лингвистических исследований РАН
Андрианова Д. В., научный сотрудник
E-mail: yakonukdar@yandex.ru

Institute of the linguistic researches
Andrianova D. V., researcher
E-mail: yakonukdar@yandex.ru