

УДК 81'33

**КОРПУСНЫЕ ТЕХНОЛОГИИ В ИЗУЧЕНИИ КОЛЛОКАЦИЙ
(НА ПРИМЕРЕ СЕРВИСОВ «ANTCONC» И «SKETCHENGINE»)**

Палийчук Д.А.

В данной работе рассматриваются возможности исследования коллокаций посредством корпусных технологий. Приведены определения понятий «корпус» и «корпусный менеджер». Рассмотрено определение «коллокации» в широком смысле и с точки зрения корпусной лингвистики. Проанализированы основные преимущества использования корпусных инструментов для изучения коллокаций. Проведено описание функционала двух наиболее популярных корпусных сервисов: «AntConc» и «SketchEngine».

Ключевые слова: корпус, корпусные технологии, корпусный менеджер, коллокация, AntConc, SketchEngine.

**CORPUS TECHNOLOGIES IN THE STUDY OF COLLOCATIONS
(BY THE EXAMPLE OF “ANTCONC”
AND “SKETCHENGINE” SERVICES)**

Palytchuk D.A.

This work reviews opportunities of studying collocations using corpus technologies. The definitions of the concepts “corpus” and “corpus manager” are given. The definitions of “collocation” in the broad sense and from the corpus linguistics point of view are considered. The main advantages of using corpus instruments for studying collocations are analysed. The description of the functionality of the two most popular corpus services – “AntConc” and “SketchEngine” has been carried out.

Keywords: corpus, corpus technology, corpus manager, collocation, AntConc, SketchEngine.

Современный этап развития лингвистической науки характеризуется активным использованием корпусов текстов в исследованиях различных

аспектов языка за счёт стремительного развития и внедрения компьютерных и информационных технологий.

Под лингвистическим корпусом текстов мы понимаем большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических или прикладных задач [5, с. 5]. С помощью корпусов становится доступным проведение ускоренного, качественного и достоверного исследования языка, что связано с автоматизацией процессов и обширным функционалом. Корпус – это не только мощный инструмент исследования языка, но и новая идеология, ориентирующая исследователя на текст как главный объект теоретической рефлексии [10, с. 14]. Работа с корпусом текстов проводится посредством корпусного менеджера (корпус-менеджера) – специализированной поисковой системы, включающей программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме [5, с. 50]. Также корпусный менеджер – это мощный инструмент для создания своего собственного корпуса текстов (подкорпуса) или же для загрузки и использования уже существующих массивов данных [4, с. 29].

В современных корпусах предлагается множество инструментов, способных удовлетворить потребности ученых для решения различных лингвистических задач. Посредством корпусных технологий можно осуществить не только точный анализ отдельных фактов реализации языковых единиц, но и выявление общих закономерностей, присущих языковой системе. Эффективное изучение лексической сочетаемости также возможно благодаря корпусным технологиям.

Термин «*коллокация*» занимает одно из центральных мест в современной лингвистике. В широком смысле коллокация – это комбинация двух и более слов, имеющих тенденцию к совместной встречаемости [6, с. 137]. В настоящее время существует множество подходов к определению данного понятия. Мы

придерживаемся определения в рамках корпусной лингвистики, или же статистического подхода, где основополагающим признаком является частота совместной встречаемости: коллокации – статистически устойчивые словосочетания. При этом статистически устойчивое сочетание может быть как фразеологизированным, так и свободным [6, с. 138].

Значительная роль коллокаций и необходимость их исследования очевидны: они встречаются во всех естественных языках; представляют главную особенность языка; коллокация конкретизирует значение слов, которые в нее входят; путь объединения слов в коллокации фундаментален для всего языкового использования; коллокация может «предсказать» окружение главного слова.

Сегодня получить информацию о частоте определенного сочетания в языке стало доступным благодаря большим репрезентативным корпусам текстов. Однако В.П. Захаров и М.В. Хохлова отмечают, что высокой величины частоты совместной встречаемости недостаточно, чтобы говорить о предпочтительной сочетаемости тех или других слов [6, с. 138]. Так были разработаны статистические меры (меры ассоциации, меры ассоциативной связанности), которые позволяют вычислить силу синтагматической связи между элементами внутри коллокации. На данный момент существует значительное количество мер ассоциации. Наиболее распространенными являются *MI*, *t-score*, *log-likelihood*, *logDice*.

– *MI* (mutual information) позволяет сравнить зависимые контекстно-связанные частоты с независимыми.

– *T-score* вычисляет частоту совместной встречаемости ключевого слова и его коллокатами.

– *Log-likelihood* – логарифмическая функция правдоподобия.

– *LogDice* – нормализованная форма меры *Dice*, которая основана только на частоте совместной встречаемости и независимых частотах и не учитывает размер корпуса (текста), в отличие от рассмотренных вышеуказанных мер [2, с. 76]. *LogDice* является более современной мерой ассоциативной связанности.

Некоторая часть корпусных менеджеров позволяет производить вычисление вышеуказанных мер.

На текущий момент корпусные технологии являются актуальным инструментом изучения коллокаций для многих исследователей [см.: 1; 3; 9; 11; 12; 13]. Существует множество корпусных программ и сервисов, предоставляющих возможность осуществить исследования коллокаций. Наиболее популярными среди исследователей в данной области являются «AntConc» и «SketchEngine».

«AntConc» – бесплатная программа, разработанная Лоуренсом Энтони и предназначенная для статистических исследований текстов. Данный сервис позволяет работать с файлами формата .txt.

В «AntConc» предлагается семь инструментов анализа текстовых данных:

1. Concordance – инструмент, демонстрирующий все контексты слова или словосочетания в тексте в формате *KWIC* (Key Word in Context).

2. Concordance Plot показывает результаты поиска в виде штрих-кода, что позволяет визуально оценить, в какой части текста встречается искомый элемент.

3. File View отображает текст отдельных выбранных файлов с маркировкой цветом указанных в поисковой строке объектов.

4. Clusters / N-grams – инструмент, осуществляющий отбор группы слов с заданным количеством элементов слева и справа от искомого слова.

5. Collocates – позволяет провести статистический анализ по словам, находящимся слева и справа от заданного элемента. Таким образом, данный инструмент дает возможность выявлять коллокации и измерять связность слов в тексте.

6. Word List предназначен для составления упорядоченного по частотности списка словоупотреблений, иными словами, частотного словаря конкретного корпуса.

7. Keyword List – инструмент, определяющий, какие слова отличаются необычно высокой или низкой частотой употребления по сравнению с эталонным корпусом.

Программа «AntConc» оснащена тремя наиболее распространенными мерами ассоциативной связанности (MI, t-score, log-likelihood), а также комбинацией мер MI и log-likelihood.

«SketchEngine» – Интернет ресурс, снабженный многочисленными инструментами корпусного анализа. Данный проект разработан корпусным лингвистом Адамом Киллгарриффом и чешским программистом Павлом Рыхли. «SketchEngine» обладает внушительным количеством языковых ресурсов более чем на 85 языках мира и предоставляет возможность создания, разметки и использования собственных корпусов. Сервис является коммерческим с предоставлением пробного периода (30 дней). «SketchEngine» позволяет работать с данными в форматах .doc, .docx, .htm, .html, .pdf, .ps, .tar.bz2, .tar.gz, .tgz, .tmx, .txt, .vert, .xml, .zip.

«SketchEngine» обладает широким набором инструментов и функций, среди которых основными являются:

1. Concordance позволяет осуществлять поиск по словоформе, лемме, словосочетанию и морфосинтаксической метке. При работе сданной функцией можно выбрать формат KWIC: отображение искомого элемента с ограничением по количеству слов слева и справа или Sentence (отображение искомого объекта в целом предложении с возможностью расширения контекста).

2. Word list – инструмент, позволяющий сформировать список слов в корпусе с указанием их частотности (как и в программе «AntConc»). В «SketchEngine» предоставляется возможность выбрать элементы (буквы, буквосочетания), которые должны входить в состав искомых объектов.

3. Keywords and Terms находит ключевые слова и термины (односложные и состоящие из нескольких компонентов) в корпусе.

4. Collocations вычисляет коллокации на основе более современной меры ассоциативной связанности logDice.

5. Word Sketch считается отличительной чертой SketchEngine среди других корпусных менеджеров. Данный инструмент осуществляет автоматическое построение коллокационных профилей или скетчей. Под скетчем понимается описание речевого поведения слова, полученное автоматически путем обобщения информации о всех контекстах, в которых исследуемое слово встретилось в корпусе [7, с. 110]. В скетчах содержится информация о сочетаемости с другими словами и о силе их связи.

6. Word Sketch Differences позволяет сравнить скетчи для двух лексических единиц.

7. Thesaurus – инструмент для автоматического создания дистрибутивного тезауруса.

8. Trends отображает изменения частоты слов в корпусе: частота каких слов возросла или сократилась, а также какие новые слова возникли [8, с. 111]. Данная функция позволяет проводить диахронический анализ лексики.

9. WebBootCaT – инструмент, позволяющий создать собственный корпус на основе текстов из сети Интернет. WebBootCaT производит автоматический поиск данных по Интернету с возможностью управления: пользователь может использовать слова-ключи в качестве исходных данных; устанавливать количество запросов; задавать длину цепочек слов; указывать минимальный и максимальный размер страницы.

Сравнив функционал «AntConc» и «SketchEngine», мы можем прийти к выводу, что данные программы позволяют эффективно работать с языковыми данными, в частности проводить исследование коллокаций. Они дают возможность устанавливать особенности поведения слова в контексте, выявлять коллокации и анализировать их особенности, получать списки наиболее встречаемых лексических сочетаний и др.

Однако, несмотря на схожий набор инструментов и функций, мы можем отметить следующие основания для отличительных черт «AntConc» и «SketchEngine»: наличие свободного доступа, технические характеристики и способы представления данных. Несмотря на отсутствие бесплатного доступа,

«SketchEngine» определяет коллокации на основе более современной меры ассоциативной связанности, позволяет работать с файлами различных форматов, а также обладает более удобным для пользователя интерфейсом. На наш взгляд, «SketchEngine» является более эффективным корпусным средством для изучения коллокаций.

Список литературы:

1. Андрианова Д.В. О некоторых возможностях выявления коллокаций с помощью интернет-технологий // Вестник Воронежского государственного университета. Серия: филология. Журналистика. 2019. № 3. С. 8-10.
2. Богоявленская Ю.В., Палийчук Д.А. Меры ассоциации для установления силы семантико-синтагматической связи элементов словосочетания // Гуманитарные исследования. История и филология. 2022. № 5. С. 69-78.
3. Горина О.Г. Применение методов корпусной лингвистики для определения контекстно-специфических слов и коллокаций // Вестник Ленинградского государственного университета им. А.С. Пушкина. 2011. Т. 7. № 3. С. 27-36.
4. Долгих З.Б. Обзор ряда корпусных возможностей в сфере лингвистических исследований (на примере анализа средств градуирования в португальском языке) // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2018. № 5 (795). С. 21-32.
5. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник для студентов, обучающихся по направлению подготовки бакалавров и магистров 035700 «Лингвистика». 2-е изд., перераб. и доп., Иркутск: СПБГУ, 2013. 144, [3] с.
6. Захаров В.П., Хохлова М.В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная

лингвистика и интеллектуальные технологии. Вып. 9 (16). М.: РГГУ, 2010. С. 137-142.

7. Котюрова И.А. Корпусные исследования с помощью сервиса AntConc в условиях работы в вузе // Язык и культура. 2020. № 52. С. 36-50.

8. Кротова Е.Б. Sketch Engine для лингвистических исследований // Германистика сегодня: материалы Международной практической конференции (Казань, 16-17 октября 2018 г.) / Под ред. М.А. Кульковой. Казань: Изд-во Казан. ун-та, 2019. С. 107-112.

9. Павельева Т.Ю. Изучение коллокаций на основе лингвистических корпусов текстов // Вестник Тамбовского университета. Серия: Гуманитарные науки. 2016. Т. 21. №. 3-4 (155-156). С. 56-61.

10. Плунгян В.А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. 2008. № 2 (16). С. 7-20.

11. Трифонова И.С., Левенкова А.Ю. Формирование коллокационной компетенции у студентов направления «Международные отношения» с использованием методов корпусной лингвистики // Вестник Томского государственного университета. 2020. № 452. С. 219-228.

12. Хохлова М.В. Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia. Серия: Инструментарий русистики: корпусные подходы. Хельсинки, 2008. № 34. С. 343-357.

13. 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. 2010. № 6. С. 30-40.

Сведения об авторе:

Палийчук Дарья Александровна – ассистент кафедры лингвистики и профессиональной коммуникации на иностранных языках Уральского

федерального университета имени первого Президента России Б.Н. Ельцина (Екатеринбург, Россия).

Data about the author:

Palytchuk Darya Alexandrovna – Assistant Professor of Department of Linguistics and Professional Communication in Foreign Languages, Ural Federal University named after the first President of Russia B.N. Yeltsin (Yekaterinburg, Russia).

E-mail: dasha.paliichuck@yandex.ru.