

УДК 004.8:81'33

КОРПУСНАЯ ЛИНГВИСТИКА: ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ИЗУЧЕНИИ ИНОСТРАННЫХ ЯЗЫКОВ

П.Д. Щетко, 1 курс

Научный руководитель – **И.И. Дегтярева**, старший преподаватель
Полесский государственный университет

В современной лингвистике одной из приоритетных задач остаётся сбор, систематизация и хранение фактического языкового материала для исследовательских целей. Решение этой задачи сегодня достигается путём создания крупных электронных коллекций текстов различной функциональной направленности. Использование компьютерных технологий и телекоммуникационных сетей позволяет не только хранить большие объёмы текстов, но и эффективно осуществлять их поиск и автоматизированную обработку [1, с. 57].

Значимость формирования электронных текстовых собраний привела к выделению этого направления в отдельный раздел прикладной лингвистики – корпусную лингвистику [1, с. 57].

Корпусная лингвистика – раздел прикладной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов с помощью компьютеров [1, с. 57].

Данное определение позволяет выделить два ключевых аспекта корпусной лингвистики:

1. создание корпусов текстов, оснащённых автоматическими инструментами их использования;
2. разработку методов экспериментального исследования различных языковых уровней на основе корпусов разных типов [1, с. 57].

Для решения практических исследовательских задач в корпусной лингвистике используются специализированные программы – корпус-менеджеры. Они позволяют создать собственный исследовательский или иллюстративный корпус путём загрузки конкретного языкового материала.

Корпусный менеджер представляет собой программу, при помощи которой производится поиск данных в корпусе и получение статистической информации в удобной для пользователя форме. Наиболее известны такие корпусные менеджеры, как WordSmith, AntConc, Sketch Engine, TermoStat [2, с. 3].

Рассмотрим функциональные возможности корпусных менеджеров на примере платформы AntConc. Для иллюстрации был сформирован корпус из 10 текстов отчётной документации Конгресса США общим объёмом 59 960 токенов. Автоматическое формирование частотного списка позволяет вывести лексические единицы от наиболее частотных к наименее частотным. При этом важно учитывать ограничения автоматического анализа: программа не может самостоятельно выделить термины, поэтому отбор единиц для исследования должен производиться исследователем вручную. Функционал корпусной лингвистики ещё развивается, и попытки автоматического выделения терминов часто могут быть ошибочными. Наилучшие результаты обнаруживаются при автоматическом выделении молексемных терминов [2, с. 3].

Для изучения контекстного употребления термина и его сочетаемости применяется функция KWIC (key word in context) – представление списка примеров в виде конкорданса с централизации-

ей по левому и правому контексту. Функция PLOT позволяет установить, в каких конкретно документах употребляется искомая единица, и её частотность. Функция Cluster позволяет установить типичные сочетания по левому или правому контексту. Функция N-Gram позволяет выявлять устойчивые словосочетания, которые могут быть перспективны для дальнейшего анализа [2, с. 4].

Особый интерес для изучения иностранного языка представляет поиск коллокаций – устойчивых сочетаний слов, которые делают речь естественной. Для оценки степени связности компонентов словосочетания используются статистические меры ассоциации, такие как t-score, MI-score и Log Dice. Данные меры позволяют автоматически выделять из корпуса устойчивые глагольные сочетания и оценивать их эффективность. Например, исследование сочетаемости глагола take проводилось на материале корпуса English Web Corpus (EWC) объёмом 52 млрд слов, который включает тексты разных жанров и тематик. Анализ проводился с помощью трёх мер ассоциации: t-score, MI-score и Log Dice. Мера взаимной информации (MI) показывает силу связи между словами: высокий MI характерен для редких, но очень устойчивых сочетаний. Мера t-score, напротив, лучше работает с частотными сочетаниями. Сбалансированная мера Log Dice, используемая в Sketch Engine, не зависит от размера корпуса. Применение этих мер позволяет изучающему язык не просто заучивать отдельные слова, а осваивать их типичное окружение, что значительно повышает естественность речи [3, с. 4].

Для решения сложных поисковых задач, выходящих за рамки простого поиска по слову, корпусные менеджеры поддерживают специальные языки запросов. Одним из таких средств являются регулярные выражения (RegEx) – мощное, гибкое и эффективное средство обработки текстов, представляющее собой миниатюрный язык программирования, предназначенный для описания и разбора текста. При поддержке со стороны конкретной программы или языка программирования регулярные выражения способны вставлять, удалять, выделять и выполнять другие операции с текстовыми данными любого вида. В контексте корпусной лингвистики регулярные выражения позволяют искать слова по шаблону, а не по конкретной словоформе: например, найти все шестибуквенные слова, начинающиеся на up, или все адреса электронной почты в тексте [4, с. 25].

Таким образом, информационные технологии играют ключевую роль в современной корпусной лингвистике, предоставляя исследователям и преподавателям иностранных языков широкий спектр инструментов для анализа языкового материала. Корпусные менеджеры, такие как AntConc, Sketch Engine и другие, позволяют не только формировать и хранить электронные коллекции текстов, но и эффективно извлекать из них статистическую и контекстуальную информацию. Функции KWIC, PLOT, Cluster и N-Gram дают возможность изучать сочетаемость лексических единиц, а статистические меры ассоциации (t-score, MI-score, Log Dice) – выявлять устойчивые коллокации, необходимые для естественной речи. Языки запросов, включая регулярные выражения, открывают доступ к сложному шаблонному поиску. В совокупности эти технологические решения позволяют изучающим иностранный язык перейти от интуитивного освоения лексики к обучению, основанному на реальных языковых данных, что значительно повышает эффективность образовательного процесса.

Список использованных источников

1. Информационные технологии в лингвистике [Электронный ресурс] // ТПУ. – Режим доступа: <https://portal.tpu.ru/SHARED/t/tatvlad/Uchebnaja/Tab4/Posobie.pdf> – Дата доступа: 15.04.2026.
2. Использование корпусных инструментов в исследовании терминологии (на материале английского политического языка) [Электронный ресурс] // Политическая лингвистика. – Режим доступа <https://politlinguistika.ru/images/4-2024/4-2024-229-234.pdf> – Дата доступа: 15.04.2026.
3. Возможности изучения сочетаемости и устойчивости лексических единиц статистическими методами (на примере глагола take) [Электронный ресурс] // КиберЛенинка. – Режим доступа: <https://cyberleninka.ru/article/n/vozmozhnosti-izucheniya-sochetaemosti-i-ustoychivosti-leksicheskikh-edinit-statisticheskimi-metodami-na-primere-glagola-take/viewer> – Дата доступа: 15.04.2026.
4. Регулярные выражения [Электронный ресурс] // O'Reilly. – Режим доступа: <https://www.pmsv.ru/wp-content/uploads/2017/03/OReilly-%D0%A0%D0%B5%D0%B3%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D1%8B%D0%B5-%D0%B2%D1%8B%D1%80%D0%B0%D0%B6%D0%B5%D0%BD%D0%B8%D1%8F-3-%D0%B5->

%D0%B8%D0%B7%D0%B4.-

%D0%94%D0%B6.%D0%A4%D1%80%D0%B8%D0%B4%D0%BB.pdf – Дата доступа: 15.04.2026.