Принципы построения лингвистического корпуса для исследовательских и образовательных целей

Окунева Ирина Олеговна
кандидат филологических наук
доцент кафедры иностранных языков
Российского государственного
гуманитарного университета (РГГУ)

Определение корпуса в корпусной лингвистике

- Корпус - это набор текстов, письменных или устных, собранных по определенному принципу, которые хранятся в электронном виде и доступны для качественного и количественного анализа.
- В прошлом этот термин в большей степени ассоциировался с совокупностью работ, например со всеми произведениями одного автора. Однако с появлением компьютеров большие объемы текстов стало возможно хранить и анализировать с помощью аналитического программного обеспечения.

Ключевое свойство корпуса

- Ключевым свойством корпуса является его репрезентативность, т.е. степень соответствия черт и свойств выбранных единиц характеристикам, свойственным всей генеральной базе данных в целом.

Принципы подбора и систематизации текстового материала для корпуса

- Цель и задачи использования корпуса
- Размер корпуса
- Жанровые особенности текстов
- Форма донесения информации (письменно, устно)
- Стили речи
- Специализированность vs общая направленность текстов
- Экстралингвистические факторы создания текстов
- Прагматическая составляющая дискурсов
- Уровень сложности языка

Виды корпусов и источники текстов

Письменные

- Сканирование текста
- Набор текста
- Загрузка текста из интернета
- Использование файлов, которые уже существуют в электронном виде, например, работы студентов

Устные

- Мультимедийные элементы
- Транскрибированные тексты устной речи

Один час записанной устной речи обычно составляет приблизительно от 12 000 до 15 000 слов.

Виды лингвистического анализа при помощи корпусов

Количественный

- Узнать частоту появления слов в текстах выборки
- Возможность сравнить частоту появлений с данными другого корпуса
- Б`ольшая частота может свидетельствовать о приоритете определенных концептов в исследуемом дискурсе

Качественный

- Получить данные о том, как определенное слово или фраза используется в корпусе
- Выявить модели сочетаемости исследуемых слов в контексте
- Составить вокабуляр, характерный для исследуемого дискурса

Корпус или словарь?

Корпус

Преимущества:

- Широкий контекст
- Множество контекстов и вариантов словоупотреблений

Недостатки:

- Необходимо самостоятельно оценивать степень общеупотребительности встречающихся словосочетаний
- Отсутствует определение значения или перевод слова

Словарь

Преимущества:

- Четкое определение или перевод слова
- Типичные примеры словоупотребления, исключающие ошибки и диалектное употребление

Недостатки:

- Риск освоить устаревшее словоупотребление
- Малочисленность и недостаточность аутентичных примеров

Использование корпусов в исследовательских и образовательных целях

- Статистический анализ частотности лексических единиц, грамматических моделей, словосочетаний
- Определение устойчивых коллокаций
- Создание тематических вокабуляров
- Создание «интеллект-карт» / «ассоциативных карт» (mind maps) культурных концептов и специальных дискурсов
- Сравнительный анализ разнообразных аспектов языка и различных языков
- Составление перечня трудностей и наиболее часто встречающихся ошибок учащихся

Благодарю за внимание