

ВОЗМОЖНОСТИ КОРПУСНОЙ ЛИНГВИСТИКИ В ИЗУЧЕНИИ РУССКОГО СЛОВООБРАЗОВАНИЯ

Аблаева Нодира

Соискатель, Университет-Маъмуна.

Джуманиязова Интизор

Соискатель, Университет-Маъмунаю.

Аннотация: Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов с использованием компьютерных технологий. Корпусный подход как метод лингвистического исследования ориентирован на прикладное изучение языка. В настоящее время все современные лингвистические исследователи и составители словарей все больше используют объёмные корпуса текстов. Далее описаны способы использования корпусов, в частности, на основе представительного объема корпуса возможность обучения русской грамматике и создания современного словообразовательного словаря русского языка. Многоаспектное применение корпусов свидетельствуют о значительной роли корпусной лингвистики и представления ее в качестве инструмента современного лингвистического исследования.

Ключевые слова: корпусная лингвистика, Национальный корпус русского языка, текст, конкорданс, автоматическая семантическая разметка, словари, словообразовательные характеристики, исследование.

Annotatsiya: Korpus lingvistikasi - kompyuter texnologiyalaridan foydalangan holda lingvistik korpuslarni qurish va ulardan foydalanishning umumiy tamoyillarini ishlab chiqish bilan shug'ullanadigan hisoblash tilshunosligi bo'limi. Korpus yondashuvi lingvistik tadqiqot usuli sifatida amaliy tilni o'rganishga qaratilgan. Hozirgi vaqtda barcha zamonaviy lingvistik tadqiqotchilar va lug'at tuzuvchilari katta matn korpuslaridan tobora ko'proq foydalanmoqda. Quyida korpusdan qanday foydalanish, xususan, korpusning vakillik hajmidan kelib chiqqan holda, rus grammatikasini o'rgatish va rus tilining zamonaviy so'z yasalish lug'atini yaratish imkoniyati tasvirlangan. Korporaning ko'p qirrali ishlatilishi korpus lingvistikasining muhim rolini va uning zamonaviy lingvistik tadqiqotlar vositasi sifatida taqdim etilishini ko'rsatadi.

Kalit so'zlar: korpus lingvistikasi, rus tilining milliy korpusi, matn, muvofiqlik, avtomatik semantik belgilar, lug'atlar, so'z yasalish xususiyatlari, tadqiqot.

Abstract: Corpus linguistics is a branch of computational linguistics that develops general principles for the construction and use of linguistic corpora using computer technology. The corpus approach as a method of linguistic research is focused on applied language learning. At present, all modern linguistic researchers and compilers of dictionaries are increasingly using voluminous text corpora. The following describes ways to use corpora, in particular, based on the representative volume of the corpus, the possibility of teaching Russian grammar and creating a modern word-building dictionary of the Russian language. The multifaceted application of corpora

testifies to the significant role of corpus linguistics and its presentation as a tool for modern linguistic research.

Keywords: corpus linguistics, Russian National Corpus, text, concordance, automatic semantic markup, dictionaries, word-formation characteristics, research.

Введение. Российский лингвист доктор филологических наук Владимир Плунгян в своих рассуждениях о новом взгляде на язык, отмечал, что «если мы хотим назвать такую область лингвистики, которая по определению является суперсовременной, то первое, что приходит в голову, — это как раз лингвистика корпусов» [Плунгян В., 2013].

Корпус языка и корпусная лингвистика – та лингвистическая сфера, которая с конца XX века начала развиваться очень стремительно. Причиной тому послужили формирующиеся современные интеллектуальные программные системы, предназначенные для обработки текстов на естественном языке. В то же время, эти системы требуют большой экспериментальной лингвистической базы. Эти два фактора – технические возможности и массивы текстов, на наш взгляд, и обусловили растущий спрос на корпусные данные. В настоящее время все современные лингвистические исследователи и составители словарей все больше используют объёмные корпуса текстов [Abdurakhmonova N., 2020]

В материалах свободной энциклопедии Википедии дано следующее определение лингвистического / языкового корпуса текстов – «это большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [<https://ru.wikipedia.org/wiki/>]

Под машиночитаемым форматом нужно понимать представление данных в формате, который может читать компьютер. Это означает, что машиночитаемые данные должны быть структурированными.

Также корпусом первого порядка называют любое собрание текстов, объединённых одним признаком, к примеру, язык автора, жанр, время создания текстов.

Основная часть. Лингвистические корпуса текстов появляются во второй половине XX века. В настоящее время корпуса созданы для многих языков мира, в частности, для венгерского, испанского, датского, итальянского, словенского, польского, чешского, эстонского, хорватского, словацкого, японского, русского. Национальный корпус русского языка (НКРЯ) охватывает почти 4.5 млн русских текстов общей длиной более 1,5 млрд слов. В настоящее время в состав НКРЯ вошли 38 корпусов: Основной корпус, Газетный корпус, Региональный корпус, Поэтический корпус, Параллельный корпус, Синтаксический корпус, Диалектный корпус, Обучающий корпус, Устный корпус, Акцентологический корпус, Мультимедийный корпус, Мультимедийный параллельный корпус, Древнерусский корпус, Церковнославянский корпус, Корпус берестяных грамот и Старорусский корпус [2]. Параллельный корпус содержит 22 параллельных подкорпуса общим объемом 165 млн слов, к примеру, английский подкорпус (43 млн), китайский (4,4 млн), башкирский (550 тыс), многоязычный (5 млн) [Национальный корпус русского языка].

Эффективность и целесообразность создания текстовых корпусов обусловлено достаточно большим объёмом данных из реальных контекстов, а также возможностью

многokrратно использовать однажды созданный корпус для решения самых разнообразных лингвистических задач.

Корпусный подход как метод лингвистического исследования ориентирован на прикладное изучение языка. Например, лексикографический анализ на базе корпусов раскрывает контекстное употребление слов, особенно синонимов, частотную сочетаемость их с другими словами, регулярность в тех или иных стилях, и четко определяет их семантику. Метод анализа на базе корпусов базируется на методах статистического и качественного анализа текста.

Уже существуют и активно задействованы многие типы корпусов текстов. Так, для изучения различных аспектов функционирования языковой системы используются исследовательские корпусы, для подтверждения и обоснования лингвистических фактов – учебные корпусы.

«В настоящее время конкордансы являются современным эффективным инструментом анализа текста, которые следует активно применять в практике преподавания языка и в лингвистических задачах» [Соснина Е.П., 2015].

Конкорданс-программа предназначена для выполнения установленной лингвистической задачи, к примеру, поиск в контексте определенной морфемы, слова или в словосочетания. Конкордансы нашли широкое применение в области прикладного языковедения, эти специальные программы представляют новые возможности для эффективного изучения языка и лексической обработки различных текстов.

В решении сложных лингвистических задач используются также статистические методики, которые особенно эффективны в машинном переводе, распознавании и синтезе речи, в системах автоматической обработки текста. Корпусы, в этом отношении, могут представить внушительную базу данных для исследований по грамматике и составлению словарей. Нужно отметить, что современные исследования по составлению словарей и грамматик в основном используют представительные корпусы текстов. «Развитие современных интеллектуальных программных систем, предназначенных для обработки текстов на естественном языке, также требует большой экспериментальной лингвистической базы. Спрос на корпусные данные совпал с появлением соответствующих технических возможностей» [Захаров В.П., 2005; 3-4].

Национальный корпус русского языка позволяет по любому слову составить конкорданс – это список всех употреблений данного слова в контексте, также в этом списке указаны ссылки на источник. На основе корпуса можно получить данные о частоте словоформ, лексем, грамматических категорий, проследить изменение частот и контекстов в различные периоды времени, получить данные о совместной встречаемости лексических единиц и т.д. Корпусы служат источником и инструментом в подготовке разнообразных исторических и современных словарей. Большой объем данных также используются для построения и уточнения грамматик и в целях обучения языку.

В НКРЯ все словоформы текстов получают автоматическую семантическую разметку, которая основана на наборе дискретных семантических характеристик, приписываемых в словаре. «В настоящее время в Корпусе реализована система поиска по лексико-семантическим признакам, основанная на частичной семантической разметке текстов. При такой разметке большинству слов в тексте приписывается один или несколько семантических и словообразовательных признаков, например, 'лицо', 'вещество',

'пространство', 'свойство человека', 'диминутив', 'отглагольное имя' и т. п.» [Национальный корпус русского языка].

«В основу семантической разметки положена система классификации русской лексики, принятая в базе данных "Лексикограф", которая разрабатывалась с 1992 г. в Отделе лингвистических исследований РАН под руководством Е. В. Падучевой и Е. В. Рахилиной... Словник семантического словаря базируется на морфологическом словаре системы DIALING, представляющим собой расширение Грамматического словаря русского языка А. А. Зализняка» [Национальный корпус русского языка].

Лексико-семантическая информация, приписываемая произвольному слову в тексте, состоит из трех групп помет:

- разряд (например, имя собственное, возвратное местоимение);
- собственно лексико-семантические характеристики (например, тематический класс лексемы, признаки каузативности, оценки);
- деривационные (словообразовательные) характеристики (например, «диминутив», «отадъективное наречие»).

Лексико-семантическая информация имеет различную структуру для разных частей речи. Кроме того, каждый из разрядов существительных - имена предметные, непредметные и собственные - имеет свою структуру помет.

Словообразовательные характеристики включают несколько типов:

- морфосемантические словообразовательные признаки (например, «диминутив», «каритив», «семельфактив»);
- разряд производящего слова (например, отглагольное существительное или отадъективное наречие);
- лексико-семантический (таксономический) тип производящего слова (например, наречие, образованное от прилагательного размера);

морфологический тип словообразования (субстантивация, сложное слово) [Национальный корпус русского языка].

Особо надо отметить, что корпуса используются лингвистами-теоретиками для проверки и доказательства своих теорий.

Компьютерные лингвисты, как особый класс пользователей, используя статистические и лингвистические закономерности в текстах, создают компьютерные модели языков. «Корпусы используются для разработки и настройки различных автоматизированных систем (машинный перевод, распознавание речи, информационный поиск)» [Захаров В.П., 2005; 7-8].

Выводы. Корпусная лингвистика – область лингвистики, в которой создаются, совершенствуются корпуса текстов, а также применяются в качестве инструмента лингвистического исследования.

Под корпусом в современной лингвистике нужно понимать совокупность текстов в электронной форме. Тексты собраны по определенным критериям, в соответствии с конкретной исследовательской задачей, и отражают ту или иную сферу языка.

Работая в корпусе данных, можно составить конкорданс - список всех употреблений слова в контексте со ссылками на источник. Представительный массив языковых данных за определенный период позволяет исследовать динамику изменения лексического состава

языка, анализировать лексико-грамматические характеристики разных жанров и у разных авторов.

Корпусы также служат инструментом в подготовке разнообразных исторических и современных словарей. Данные корпусов используются прикладными лингвистами (преподавателями, переводчиками) для построения грамматик, в обучении языку и для решения своих профессиональных задач. Корпусная лингвистика имеет практические механизмы создания и использования представительных массивов языковых данных, которые предназначены для лингвистических исследований в интересах широкого круга пользователей.

Таким образом, можно сказать, что посредством компьютерных технологий корпус превратился из метода работы с языковым материалом в эффективный инструмент исследования.

Список использованной литературы:

1. Abdurakhmonova N. Komputer lingvistikasi. Учебник – Globe edit publishing, 2020, 395b.
2. Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. – СПб., 2005. – 48 с.
3. Национальный корпус русского языка. 2003-2022. Доступен по адресу: ruscorpora.ru
4. Плунгян Владимир. Корпусная лингвистика. 14.01.2013. <https://postnauka.ru/video/7783>.
5. Соснина Е.П. Параллельные корпуса в обучении языку и переводу. - 2015. http://ling.ulstu.ru/linguistics/resourses/literature/articles/corpus_education_translation/
6. <https://ru.wikipedia.org/wiki/>