КОРПУСНАЯ ЛИНГВИСТИКА: НОВЫЕ ПОДХОДЫ К АНАЛИЗУ ЯЗЫКА И ИХ ПРИЛОЖЕНИЯ В ОБУЧЕНИИ ИНОСТРАННЫМ ЯЗЫКАМ

Сухроб Собирович Авезов
Преподаватель кафедры русского литературоведения
Бухарский государственный университет
1990senigama@gmail.com
https://doi.org/10.5281/zenodo.8157748

Аннотация. Статья освещает применение корпусной лингвистики в обучении иностранным языкам и лингвистической педагогике. Автор подчеркивает значение корпусов письменных и устных текстов в формировании списков активного словарного запаса студентов и частотных перечней терминов для профессиональных курсов. В работе также отмечается, что разработчики академических словарей и учебных материалов опираются на эти корпусы, отражающие реальное функционирование конкретного языка. Несмотря на широкое распространение и использование корпусной лингвистики в практической и научной деятельности, в статье поднимается вопрос об отсутствии полноценной теоретической базы данного подхода. Автор выражает надежду на то, что корпусная лингвистика в ближайшем будущем приобретет статус самостоятельной научной дисциплины.

Ключевые слова: Корпусная лингвистика, обучение иностранным языкам, лингвистическая педагогика, корпусы текстов, активный словарный запас, теоретическая база.

В современную эпоху лингвистическая дисциплина, известная как корпусная лингвистика, подверглась значимым трансформациям в контексте своего прогресса. Исходя из специфического подхода, традиционно применяемого в англоязычной лингвистической области и, в частности, в исследовании английской грамматики, корпусная лингвистика начала активно расширять свой сферический охват.

В соответствии с обзором С.Л. Мишлановой, история развития корпусных исследований обозначает три основных этапа: 1960-1980 гг., когда преобладали индивидуальные корпусные исследования «специфического прикладного свойства», предназначенные в большей степени для словарного составления. Второй период (1980-1990 гг.) характеризуется формированием корпусного подхода, включающего методику организации текстов в корпус, аналитические алгоритмы и связанную с этим научную методологию. Третий период, наступивший с началом XXI века и продолжающийся в настоящее время, отмечается выделением корпусной лингвистики в самостоятельную прикладную лингвистическую область, формированием метода корпусного анализа и его широким применением во всех отраслях лингвистики и связанных дисциплин. [1]

Что представляет собой корпусная лингвистика? Это дисциплина, которая кардинально отличается от большинства других лингвистических областей, поскольку она не фокусируется на прямом изучении конкретного аспекта языка. Это область,

ориентированная на систему методов для исследования языка. В настоящее время эти методы продолжают свое совершенствование, их определение до сих пор не является строго очерченным, хотя некоторые из них, такие как конкорданс, уже хорошо известны и считаются ключевыми в данном подходе.

Учитывая эти методы, мы можем применять корпусно-ориентированный подход к множеству лингвистических областей. Параллельно, корпусная лингвистика может переосмыслить нашу концептуальную базу в изучении языка, обеспечивая уточнение, детализацию и переопределение различных языковых теорий, а также предоставляя нам возможность применять теории языка, изучение которых было сложным из-за недостатка корпусов подходящего объема и компьютерных ресурсов, обладающих адекватной мощностью для их использования. [2]

Развитие корпусной лингвистики также способствовало появлению и исследованию новых языковых теорий. Она фокусируется на наборе методов для изучения языка; это не единый, консолидированный набор методов для изучения языка. Хотя возможно сделать обобщения, характеризующие большую часть того, что называется «корпусной лингвистикой», важно осознавать, что корпусная лингвистика является разнообразной сферой. [3] В рамках корпусной лингвистики существуют различия, которые разделяют и дифференцируют различные подходы к использованию корпусных данных. Однако давайте начнем с обобщений.

В начале можно определить корпусную лингвистику как «коллекцию текстов, которые могут быть обработаны машинами и считаются достаточно репрезентативными для определенных научных исследований» [4]. Объемы такого рода данных недоступны для ручного анализа, что подтверждает необходимость использования машинообработки текстов.

Второй момент заключается в том, что корпусы текстов систематически изучаются с помощью множества инструментов, которые предоставляют возможность пользователям выполнять исследования быстро и эффективно. Некоторые из этих инструментов, такие как конкордансеры, позволяют исследователям увидеть слова в контексте их употребления. Большинство этих инструментов также обеспечивают создание частотных данных, как, например, список наиболее употребляемых слов, где указываются все слова, встречающиеся в корпусе, с учетом количества их употреблений. Конкордансы и частотные данные представляют собой два типа анализа: качественный и количественный, которые имеют равную значимость для корпусной лингвистики.

Вес наших выводов, основанных на корпусе, будь то качественного или количественного характера, зависит от другого общего параметра, применимого ко всем типам корпусной лингвистики: «корпусные данные, выбранные для нашего исследования, должны точно соответствовать проблеме нашего исследования» [5]. Например, нелогично было бы изучать систему классификации существительных на языке суахили, используя корпус английских газетных статей. Мы не можем (или можем только с определенной осторожностью) формулировать общие утверждения о характере данного языка, основываясь на корпусе, который содержит только один тип текста или ограниченное число типов текстов.

А также, мы должны принять во внимание факт, что тексты внутри корпуса, которые мы считаем однородными, могут, на самом деле, быть различными. Например,

сборник статей из одной и той же газеты, даже за один и тот же день, может включать в себя заметные различия – стилистика и словарь в спортивной колонке будет отличаться от колонки международных новостей. Пользователи корпуса должны быть в курсе этих внутренних различий, поэтому исследователи иногда прибегают к статистическим методам для изучения степени вариативности в данном корпусе перед его использованием. Степень гомогенности корпуса является еще одним фактором при определении соответствия корпуса конкретным научным вопросам.

Стоит уточнить, что в данном контексте под «текстом» мы подразумеваем «файл, доступный для машинного чтения». В рамках корпусной лингвистики каждый такой файл может, например, представлять газетную статью или орфографическую транскрипцию устной речи. Однако файлы внутри корпуса не обязательно должны быть текстовыми, и в наше время уже есть примеры использования видеофайлов в качестве «текстов» для корпусов.

Все вышеперечисленное служит характеристикой того, что мы можем определить как корпусную лингвистику. Электронные корпусы обеспечивают богатый лингвистический материал для образовательных и научных целей. В современном интернет-пространстве представлено множество электронных корпусов на иностранных языках, включая такие известные, как British National Corpus (BNC), International Corpus of English (ICE), Американский национальный корпус английского языка, а также немецкоязычные корпуса LIMAS, COSMAS.

Относительно типологии текстовых корпусов в прикладной лингвистике возможно применение следующих типов:

1. Исследовательские - предназначенные для исследования различных аспектов функционирования языковой системы;

2. Иллюстративные, включая учебные (Learner Corpus) - для подтверждения и обоснования лингвистических фактов;

3. Мониторные - для исследования динамики языкового материала, проведения анализа содержания, например, корпус публицистики;

4. Статические - для исследования стилей, например, авторские корпусы или корпусы текстов писателей;

5. Мультимедийные - сочетающие текст, видео и аудио;

6. Корпусы параллельных текстов - для сопоставительного анализа текстов «оригинал-перевод» для обучения методам и приемам перевода. Есть два основных способа организации таких корпусов: «оригинал-перевод/ы» (Unidirectional), «оригинал - перевод - обратный перевод» (Bidirectional or reciprocal), упорядоченные параллельно.

7. Многоязычные (multilingual)

8. Исторические или диахронические (корпусы, позволяющие изучить эволюцию языка или определенные изменения, происходящие в языке в течение определенного периода времени). Примером может служить Хельсинкский корпус, включающий 1.5 миллиона слов из английских текстов.

Центральными методиками в области корпусной лингвистики являются: конкорданс, коллокация, анализ на основе ключевых слов и аннотация. Рассмотрим каждый из них более подробно.

1. Конкорданс или гармонизация представляет собой «инвентарь каждого вхождения слова (или образа) в текст или корпус, включая слова, которые его окружают». [6] Конкорданс, в сущности, является методом визуализации данных. Поисковый запрос и ассоциированный с ним текст организованы таким образом, чтобы была возможность оценить текстовую обстановку и визуальные связи, связанные с поисковым запросом. Часто конкорданс специфического поискового запроса в корпусе приводит к слишком большому количеству результатов для лингвиста для их чтения и анализа. В таком случае можно отобрать меньшее число примеров. Конкорданс - это полезный инструмент для исследования корпусов, но его применение ограничено способностью наблюдателя обрабатывать информацию.

2. Коллокация - в корпусной лингвистике подразумевается под коллокацией «последовательность слов или терминов, совместное появление которых в корпусе превышает ожидаемую вероятность их совместного появления». Статистическое измерение коллокаций является более достоверным методом исследования в корпусной лингвистике.

3. Анализ на основе ключевых слов. Ключевые слова - это наиболее часто употребляемые слова в сравнении с некоторым стандартом. Сопоставление частотных списков для двух корпусов может предоставить ценную информацию о различиях между этими двумя текстами. Анализ на основе ключевых слов может использоваться для анализа стиля и поиска текстов.

4. Аннотация - это обобщенное наименование для меток и синтаксического анализа, а также используется для описания других типов категоризации, которые могут быть выполнены в корпусе (например, аннотация устного корпуса для просодических особенностей, аннотация учебного английского языка для типов ошибок, аннотации анафоры и семантическая аннотация).

В современном мире корпусы письменных и устных текстов эффективно используются в обучении иностранным языкам и в области лингвистической педагогики. Базируясь на корпусах, формируются списки активного словарного запаса учащихся, частотные перечни терминов для применения в профессиональных курсах и так далее. Создатели академических словарей и учебных материалов опираются на автентичные массивы текстов. К тому же, коллекции, библиотеки и массивы текстов отражают реальное функционирование определенного языка, а их перенос в цифровую среду только стимулировал их практическое и обширное использование в прикладной лингвистике. Несмотря на то что корпусная лингвистика становится все более популярной, ее методы все чаще применяются в лингвистических исследованиях и существует большое количество курсов повышения квалификации, цель которых - научить использованию корпусов, проведению исследований на основе корпусов (например, Корпусная лингвистика: метод, анализ, интерпретация (Университет Ланкастера)), теоретическая основа корпусной лингвистики все еще не полностью разработана. Именно по этой причине ученые до сих пор не могут дать определенный ответ на вопрос: «Что такое корпусная лингвистика: новая научная дисциплина или просто информационный ресурс?» Мы ожидаем, что в ближайшем будущем ответ на этот вопрос будет найден и корпусная лингвистика станет самостоятельной научной дисциплиной.

Список использованной литературы:

1. Nigmatova L. K. Language and cultural issues in uzbek vocabulary //Scientific reports of Bukhara State University. – 2021. – Т. 5. – №. 1. – С. 30-49.
2. Sobirovich A. S. Development of a Parallel Corpus of the Uzbek and Russian Languages //Vital Annex: International Journal of Novel Research in Advanced Sciences. – 2022. – Т. 1. – №. 5. – С. 152-155.
3. Авезов С. С., Маринина Ю. А. Электронные Корпусы: Инновационный Подход К Обучению Переводу //Periodica Journal of Modern Philosophy, Social Sciences and Humanities. – 2023. – Т. 16. – С. 7-13.
4. Шарипов С. С. ЛЕКСИКОГРАФИЯ (ТАРЖИМА ЛЕКСИКОГРАФИЯСИ) РИВОЖЛАНИШИНИНГ АСОСИЙ ЖИҲАТЛАРИ //МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИСКУССТВО СЛОВА. – 2022. – Т. 5. – №. 3.
5. Nigmatova L., Avezov S. ПРИМЕНЕНИЕ МЕТОДОВ NLP В КОРПУСНЫХ ИССЛЕДОВАНИЯХ: ОСОБЕННОСТИ И ОГРАНИЧЕНИЯ //«УЗБЕКСКИЕ НАЦИОНАЛЬНЫЕ ОБРАЗОВАТЕЛЬНЫЕ ЗДАНИЯ ТЕОРЕТИЧЕСКОЕ И ПРАКТИЧЕСКОЕ СОЗДАНИЕ ВОПРОСЫ" Международная научно-практическая конференция. – 2023. – Т. 2. – №. 2.
6. Hunston S. Pattern grammar, language teaching, and linguistic variation //Using corpora to explore linguistic variation. – 2002. – С. 167-183.