ИРГИЗОВА К. В.

КОРПУСНАЯ ЛИНГВИСТИКА В ОТЕЧЕСТВЕННОМ И ЗАРУБЕЖНОМ ЯЗЫКОЗНАНИИ НА СОВРЕМЕННОМ ЭТАПЕ

Аннотация. В статье представлен краткий обзор ключевых положений корпусной лингвистики в рамках отечественного и зарубежного языкознания. Автор описывает сходства и различия между подходами к изучению корпусов в разных странах, а также говорит о перспективах их использования в различных областях современной науки о языке.

Ключевые слова: корпусная лингвистика, языковой корпус, прямые и косвенные применения корпусов, корпусный подход.

IRGIZOVA K. V.

CURRENT STATE OF RUSSIAN AND INTERNATIONAL CORPUS LINGUISTICS

Abstract. The article provides an overview of the key concepts of Russian and international corpus linguistics. The author describes the similarities and differences between approaches to the study of corpora in different countries, and also considers the prospects that corpus linguistics can facilitate in various domains of modern linguistics.

Keywords: corpus linguistics, language corpus, direct and indirect corpus applications, corpus-based approach.

Изучение языковых корпусов, начавшееся в середине XX века, привело к формированию такого направления науки о языке как корпусная лингвистика. В этой связи В. А. Плунгян, академик РАН, заведующий кафедрой теоретической и прикладной лингвистики МГУ им. М. В. Ломоносова, описывает это направление как «стремительное» и «суперсовременное» [9, с. 9]. Корпусная лингвистика обладает большим исследовательским потенциалом, однако наблюдается различие подходов к использованию корпусов в России и за рубежом.

Рассматривая явление корпусной лингвистики в российском и зарубежном научном онлайн-пространстве на базе таких платформ как Elibrary, Research Gate и Google Scholar можно обнаружить, что языковые корпусы рассматриваются:

- как одно из направлений общего языкознания в рамках корпусной лингвистики как науки;
- как один из элементов стратегии обучения профессионально-ориентированному иностранному языку;
- как база для проведения эмпирических исследований в области лексикографии и теории перевода.

Как видно из нашего определения, корпусная лингвистика применима в различных областях знаний. Однако, рассматривая данную науку с точки зрения ее популяризации, можно сказать, что Запад обращается к корпусному инструментарию гораздо чаще, чем наши соотечественники. Так, например, при русскоязычном запросе «корпусная лингвистика» платформа Elibrary выдает результаты всего по 2 323 статьям. Хотя если изменить запрос на англоязычный аналог «corpus linguistics», то количество статей возрастет до 8 517, то есть почти в 4 раза. В рамках платформы Google Scholar наблюдается похожая ситуация с разницей лишь в цифрах: по запросу «корпусная лингвистика» выходит 15 400 упоминаний, а по запросу «corpus linguistics» – в 69 раз больше, то есть более 1 000 000 совпадений. Таким образом, мы видим, что количество отечественных корпусных исследований меньше зарубежных. В нашей стране существуют не так много отдельных исследований, посвященных изучению применения новейших технологий в языкознании. Например, исследования ведутся учеными НИЛ «Интеллектуальные технологии управления текстами» ФГАУ ВО «Казанский Федеральный Университет». Авторами статей на заданную тему являются, например, А.С. Кисельников, Е.В. Харкова, О.С. Сафонкина, Е. В. Варламова [13, с. 200].

В. П. Захаров в своей книге «Корпусная лингвистика» отмечает, что в рамках данной научной дисциплины ключевым элементом является языковой корпус, под которым понимается «многогранное собрание естественных случаев употребления языка в виде текстов разной жанровой и стилистической направленности и хранящееся в электронном формате» [4, с. 8]. Его основное назначение состоит в обеспечении получения достоверной информации об употреблении слова и нахождении лексических единиц и грамматических конструкций благодаря лингвистической разметке.

Нужно сказать, что наука о корпусах на Западе начала развиваться на 40 лет раньше, чем в России. Ее реализация происходит по аналогичным принципам и стратегиям, однако существуют некоторые критерии для дифференциации. Прежде всего, это разница в практике конструирования лингвистических корпусов в плане их количества и разнообразия. Западные языковеды имеют богатый опыт в создании языковых корпусов, в частности, первый известный электронный корпус – «Брауновский корпус» Г. Кучера и Н. Фрэнсиса – был создан в середине XX века. Через несколько лет начали издаваться первые словари на базе корпусов, как например, широко известная линейка Collins' COBUILD (Collins Birmingham University International Language Database), которые представлена отдельными словарями, посвященными различным языковым элементам: фразеологизмам, метафорам, предлогам, омофонам, квантификаторам и т.д. Развитие корпусных технологий, разработка новых классификаций и параметров позволили лингвистам создавать корпусы вариантов

английского языка, таких как Wellington (новозеландский английский), Kolhapur (индийский английский) и т.д. [2, с. 46].

Отечественное языкознание несколько отстает от своих зарубежных коллег в области конструирования корпусов. Первым, и на сегодняшний день единственным завершенным корпусом русскоязычных текстов, является Уппсальский корпус текстов. Он был создан в конце XX века и в настоящее время используется мало. В связи с тем, что данный корпус не способен полностью отвечать современным требованиям в силу ограниченного объема и отсутствия лингвистической разметки (морфологических, синтаксических, семантических маркеров), начиная с 2003 года, российские лингвисты начали работу над новым проектом, который носит название «Национального корпуса русского языка» (НКРЯ), которая в настоящее время активно ведется [5, с. 85].

Следует отметить, что большинство лингвистических исследований все же проводятся на базе текстов из корпусов английского языка. Возможно, причина этого заключается в том, что период расцвета британской лингвистики приходится на 60-80-е гг. XX века, т.е. на момент создания первого корпуса английского языка. Ко всему прочему, активное развитие информационных технологий в Великобритании и США могло стать катализатором прогресса создания электронных корпусов [2, с. 44]. Одними из наиболее известных англоязычных корпусов принято считать:

- Британский национальный корпус (BNC, British National Corpus). BNC является, по сути, моделью для создания всех современных национальных корпусов. Данный корпус включает в себя более 100 млн. слов и снабжен метатекстовой и морфологической разметкой. В соотношении между письменной и устной речью, представленной текстами из СМИ, школьными ученическими работами, научными статьями и иными текстами разных жанров, можно увидеть большой разрыв – 90% к 10%. BNC отражает фактическое состояние британского английского начала нынешнего столетия. Поиск лексико-грамматических конструкций (словосочетаний, словоформ и т.д.) осуществляется с помощью корпусного менеджера XAIRA, который также позволяет найти информацию об источниках примеров текстов и данные о частоте употребления тех или иных коллокаций. В онлайн-режиме возможно осуществить только ограниченный доступ к данному корпусу (50 случайных примеров в выдаче результатов), поскольку его полная версия, предоставляемая на DVD, – платная.
- Национальный корпус американского английского (NAC, National American Corpus) был создан в качестве аналога BNC. На сегодняшний день он включает в себя 22 млн. слов. Как и в случае с большинством корпусов, полная версия NAC является платной,

однако 68% слов находится в свободном онлайн-доступе. В отличии от BNC в данном корпусе отсутствует поисковый интерфейс, таким образом поиск информации осуществляется посредством применения универсальных корпусных менеджеров, которые не ориентированы на работу с одним конкретным корпусов. Корпус обеспечен метатекстовой, частеречной и частичной синтаксической разметкой. Также в составе NAC есть так называемая разметка именованных сущностей (Named Entities), включающая в себя имена собственные, названия организаций и географических объектов [15].

Если говорить о структуре языкового корпуса, то можно отметить, что она напрямую связана с его функциональностью, а также областью и целями применения. Следовательно, для изучения и анализа определенной языковой подкатегории (синтаксиса, стилистических особенностей и т.д.) создатели корпусов должны собрать максимально возможную коллекцию текстов, относящихся к этой подкатегории. Таким образом, мы условно можем назвать корпус «репрезентативной уменьшенной моделью языка или подъязыка» [4, с. 18]. Репрезентативность корпуса понимается как его способность отражать все свойства проблемной области и выражается в определенной статистической оценке их количества. Именно эта характеристика корпуса помогает определить достоверность фактов, полученных из него. Исходя из этого, по критерию репрезентативности, следовательно, и по типу структурного наполнения, корпусы делятся на:

- Корпусы 1 типа, которые являются всеохватывающими и представляют все многообразие речевой деятельности. На настоящий момент не существует корпусов 1 типа, представленных в чистом виде, поскольку язык является настолько многогранным явлением, что рассчитать при помощи математических методов абсолютно все его свойства и категории представляется невозможным. Даже национальные общеязыковые корпусы не могу включить в себя все множество употреблений языка, однако на этапе конструирования данная категория корпусов выглядит максимально репрезентативной в сравнении с другими типами корпусов. В качестве примера все же можно привести «Брауновский корпус», который по меркам своего времени отличался достаточной репрезентативностью. В своей структуре данный корпус имел до 15 стилевых регистров, каждый из которых был представлен результатами (текстами) 80 и более выборок. Среди различных жанров в данном корпусе были представлены образцы художественной литературы, тексты научно-популярной, биографической, религиозной тематики, правительственные документы, репортажи СМИ и т.д.

- Корпусы 2 типа, которые создаются для специальных целей, чаще всего относятся к определенному типу дискурса и отражают различные лингвокультурологические феномены в процессе коммуникации. Критерием репрезентативности в данном случае является максимально возможное объективное представление какого-либо явления, интересующее создателей данного корпуса. Так, например, корпус англоязычных пословиц, отражающий использование в речи носителей языка определенного времени и географического региона, не будет релевантным при изучении английской политической метафоры [10, с. 126].

Немаловажным является и отношение лингвистов к корпусному подходу как самостоятельной науке. В этом плане взгляды отечественных ученых и их зарубежных коллег схожи. Общеизвестен тот факт, что на момент создания первых корпусов в лингвистике за рубежом активно развивался генеративный подход. Его основатель, Н. Хомский, говорил о том, что базовая информация об устройстве синтаксиса заложена в человеческом сознании с рождения, и потому, может быть применима при освоении совершенно любого языка. По его мнению, в изучении языка главным компонентом является человеческая интуиция, и, соответственно, неправильных речевых конструкций априори не существует. Данная теория подвергалась массовому обсуждению и критике. К концу XX века западные лингвисты сделали вывод о том, что создание релевантного словаря и грамматики возможно только на основе репрезентативного собрания текстов с множеством примеров актуального использования языка. Такой же концепции придерживаются и российские ученые, в частности один из руководителей проекта создания «Национального корпуса русского языка» В. Н. Плунгян. Он говорит, что корпус необходим исследователям, занимающимся систематизацией фактов об анализируемом языке, а также в академических целях, поскольку таким образом процесс освоения языковых компетенций происходит быстрее [9, с. 11].

Несмотря на то, что на сегодняшний момент корпусная лингвистика является не до конца изученной областью знаний в рамках отечественного языкознания, интерес российского научного общества к ее исследованиям возрастает с каждым годом, поскольку она создает многообещающие перспективы в сфере лингвистики. Во-первых, это новый взгляд на дискурс как реальный, а не фиктивный элемент коммуникации. Идеология корпусной лингвистики построена на том, что при работе используются не искусственно созданные тексты, а примеры живого использования языка.

Во-вторых, это акцент на количественный анализ языка, а именно исследование элементов, наиболее часто используемых в речи [6, с. 84].

В-третьих, это работа с языком в рамках синхронического и диахронического подходов.

Языковые корпуса можно применять в различных научных областях, таких как:

- Лексикография. По принципу корпуса создается большое количество не только бумажных, но и онлайн-словарей, например, словарь ReversoContext, представляющий особую ценность для переводчиков и преподавателей иностранных языков. В данном словаре значение лексической единицы распознается в контексте методом соотнесения двух текстов на разных языках. В отличие от обычного словаря, в составе которого находятся уже зафиксированные языковые нормы, корпус может дать информацию о настоящем статусе того или иного слова и его функционирования в речи;
- Лингводидактика. В сфере преподавания иностранных языков корпусы находят свое отражение в качестве наиболее релевантных источниках языкового материала, который, к тому же, можно постоянно обновлять по мере модернизации самих корпусов. Необходимость использования корпусов на занятиях по языку также обусловлена тем, что информация, полученные из них (например, частотность употребления определенных лексических и грамматических явлений), помогают при определении содержания обучению;
- Переводоведение. При обучении переводу параллельные корпуса позволяют увидеть определенные закономерности и лингвистические законы в тексте оригинала и текста перевода. В свою очередь, объектом сопоставительных корпусов является одинаковая коммуникативная направленность в двух разноязычных текстах;
- Изучение лексико-грамматического строя языка. В рамках данного аспекта корпусы дают возможность отследить появление неологизмов, сочетаемость тех или иных грамматических конструкций, процессы деноминации и т.д.

В большинстве случаев популярность использования корпусов объясняется возможностью исследования языковых закономерностей на материале большой базы текстов, обработанной и представленной в виде электронной платформы. Причем, корпус, как правило, не является аналогом стандартного электронного библиотечного каталога, так как позволяет искать отдельные фрагменты текстов по специальным параметрам и критериям, выделенным исследователем. Структура корпуса дает возможность рассматривать язык с разных сторон, выделяя определенные закономерности и формулируя новые лингвистические законы. Стоит также сказать, что корпусным исследованиям свойственна оценка речевых образцов в контексте реального применения языка. Кроме того, массив языковых данных, построенный по принципам корпусной лингвистики, может

использоваться не один раз, что подтверждает его представительность, возможность повторных лингвистических исследований и верификации их результатов [11, с. 64].

Актуальность развития корпусной лингвистики как нового направления в науке не вызывает сомнений и не оспаривается современными учеными. Возникнувшая относительно недавно, она является результатом синтеза знаний в различных областях языкознания. Так, например, в рамках сравнительно-исторического языкознания корпусная лингвистика использует технологии по реконструкции древних языков для лингвистического анализа. Так же корпусы текстов могут использоваться в качестве эмпирического и иллюстративного материала для различных лексико-грамматических явлений. Социолингвистика обращается к использованию корпусных критериев для создания пособий по изучению разновидностей языка (диалект, социолект и т.д.) [8, с. 42]. Кроме того, корпусная лингвистика позволяет найти точки соприкосновения между гуманитарными и техническими науками.

Таким образом, создание корпусов стало революцией в области анализа дискурса, позволяя делать бесчисленное множество операций с текстом за секунды, как например, разбиение фрагментов текста по нужным критериям, их маркировка и разметка. В этой связи корпус позволяет объективно и оперативно рассмотреть язык таким, какой он есть на самом деле на актуальных и «живых» примерах.

ЛИТЕРАТУРА

1. Вадяев С. Е. Электронная лексикография и корпусная лингвистика // Аспекты становления и функционирования западногерманских языков. – Самара, 2003. – С. 83–92.

2. Волоснова Ю. А. Корпусная лингвистика: проблемы и перспективы // Вестник Московского государственного университета леса – Лестной вестник. – Москва, 2006. – № 07. – С. 43–49.

3. Грудева Е. В. Как корпусная лингвистика изменила наши представления о языке // Материалы XIII выездной школы-семинара «Проблемы порождения и восприятия речи». – Череповец, 2015. – С. 108–115.

4. Захаров В. П., Богданова С. Ю. Корпусная лингвистика: учебник для студентов гуманитарных вузов. – Иркутск: ИГЛУ, 2011. – 161 с.

5. Изотов А.И. Новые направления славянского языкознания: корпусная лингвистика // Язык, сознание, коммуникация. – Москва, 2015. – С. 82–93.

6. Карамнов А.С. Количественная оценка повторяемости и сложности лексики в корпусе учебника английского // Филологические науки. Вопросы теории и практики. – 2014. – № 06 (36). – С. 82–86.

7. Куренко К. Н. Корпусная лингвистика в переводоведении // Иностранные языки: лингвистические и методические аспекты. – Тверь, 2017. – № 37 (36). – С. 318–323.

8. Майорова А. Д. Корпусная лингвистика: исторический и лингводидактический аспекты // Международный научно-исследовательский журнал. – Екатеринбург, 2017. – № 05 (59). – С. 42–46.

9. Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. – Москва, 2008. – № 02 (16). – С. 7–20.

10. Рыков В. В. Корпус текстов как реализация объектно-ориентированной парадигмы // Труды международного семинара «Диалог-2002». – М.: Наука, 2002.– С. 124–129.

11. Савчук С.О. Национальный корпус русского языка: перспективы использования в лингвистических исследованиях и в преподавании // Вестник Азиатско-Тихоокеанской ассоциации преподавателей русского языка и литературы. – Владивосток, 2011. – № 02 (03). – С. 62–67.

12. Садовникова О. Э. Прямое и косвенное использование корпусов в зарубежной лингводидактике // Научно-педагогический журнал Восточной Сибири «Magister Dixit». – Иркутск, 2013. – № 02. – С. 152–161.

13. Сафонкина О. С. Использование современных компьютерных технологий при обучении иностранным языкам // Материалы конференции «Иностранные языки в диалоге культур: экономика, политика, образование». – Саранск, 2005. – С. 200–201.

14. Ghsoon R. English Coursebooks: Prototype Texts and Basic Vocabulary Norms // ELT Joural. – 2013. – Vol. 57 (3) – pp. 260–268

15. Studiorum: Образовательный портал Национального корпуса русского языка [Электронный ресурс]. – Режим доступа: https://studiorum-ruscorpora.ru/current/ (дата обращения: 13.05.2019).