

**ВНЕДРЕНИЕ И ИССЛЕДОВАНИЕ КОРПУСНОЙ ЛИНГВИСТИКИ.
KORPUS TILSHUNOSLIGINI TADBIQ ETISH VA TADQIQ QILISH.
IMPLEMENTATION AND RESEARCH OF CORPUS LINGUISTICS.**

Рузметова Озода Рузметовна

Преподаватель кафедры языка и литературы

azadaruzmetova3@gmail.com

+998914372550

Университет Мамун

Ro'zmetova Ozoda Ro'zmetovna

Til va adabiyot kafedrası o'qituvchisi

azadaruzmetova3@gmail.com

+998914372550

Mamun universiteti

Ruzmetova Ozoda Ruzmetovna

Lecturer of the Department of Language and Literature

azadaruzmetova3@gmail.com

+998914372550

Mamun University

Аннотация. Проведено всестороннее исследование этапов формирования и развития корпусной лингвистики. Основная цель статьи заключается в анализе научных подходов к значимости данной лингвистической дисциплины, а также в

выявлении понятий и критериев, которые лежат в основе этого направления. Корпусная лингвистика является одной из наиболее перспективных и стремительно развивающихся областей языковых исследований. Лингвистика XIX века акцентировала внимание на изучении языка как самостоятельного объекта, в то время как современная лингвистика XXI века рассматривает важность исследования не в выделении абсолютных лингвистических категорий, а в практике использования лингвистических знаний. Актуальность данной статьи обусловлена тем, что лингвистические корпуса содержат огромный потенциал, который еще не полностью исследован научным сообществом. Текст, как основной объект корпусной лингвистики, в своих различных проявлениях является одной из ключевых составляющих системы языка и речевой деятельности современного носителя языка. Разнообразие содержания и объема лингвистических корпусов позволяет получать точные данные о фактическом использовании тех или иных терминов. Корпус становится важным инструментом для анализа функционирования этих терминов как в области лингвистики (морфология, синтаксис, лексика), так и в теории и практике перевода, позволяя определить стиль их формального или неформального использования.

Ключевые слова: корпусная лингвистика, языковой корпус, методология, репрезентативность, классификация, критерии

Izoh. Korpus tilshunosligining shakllanish va rivojlanish bosqichlarini har tomonlama o'rganish amalga oshirildi. Maqolaning asosiy maqsadi ushbu tilshunoslik fanining ahamiyatiga ilmiy yondashuvlarni tahlil qilish, shuningdek, ushbu yo'nalish asosidagi tushunchalar va mezonlarni aniqlashdir. Korpus lingvistikasi til tadqiqotining eng istiqbolli va jadal rivojlanayotgan sohalaridan biridir. 19-asr tilshunosligi tilni mustaqil ob'yekt sifatida o'rganishga asosiy e'tiborni qaratgan bo'lsa, XXI asrning zamonaviy tilshunosligi mutlaq lingvistik kategoriyalarni aniqlashda emas, balki lingvistik bilimlardan foydalanish amaliyotida tadqiqotning muhimligini ko'rib chiqadi. Ushbu

maqolaning dolzarbligi lingvistik korpus hali ilmiy jamoatchilik tomonidan to'liq o'rganilmagan ulkan salohiyatga ega ekanligi bilan bog'liq. Matn korpus tilshunosligining asosiy ob'ekti sifatida o'zining turli ko'rinishlarida zamonaviy ona tilida so'zlashuvchining til tizimi va nutq faoliyatining asosiy tarkibiy qismlaridan biridir. Lingvistik korpusning mazmuni va hajmining xilma-xilligi muayyan atamalarning amalda qo'llanilishi to'g'risida aniq ma'lumotlarni olish imkonini beradi. Korpus tilshunoslik (morfologiya, sintaksis, lug'at) sohasida ham, tarjima nazariyasi va amaliyotida ham ushbu atamalarning ishlashini tahlil qilish uchun muhim vosita bo'lib, ularning rasmiy yoki norasmiy qo'llanilishi uslubini aniqlash imkonini beradi.

Kalit soʻzlar: korpus lingvistikasi, til korpusi, metodologiya, reprezentativlik, tasnif, mezonlar

Abstract. A comprehensive study of the stages of formation and development of corpus linguistics has been conducted. The main objective of the article is to analyze scientific approaches to the significance of this linguistic discipline, as well as to identify the concepts and criteria that underlie this direction. Corpus linguistics is one of the most promising and rapidly developing areas of language research. Linguistics of the 19th century focused on the study of language as an independent object, while modern linguistics of the 21st century considers the importance of research not in identifying absolute linguistic categories, but in the practice of using linguistic knowledge. The relevance of this article is due to the fact that linguistic corpora contain enormous potential that has not yet been fully explored by the scientific community. Text, as the main object of corpus linguistics, in its various manifestations is one of the key components of the language system and speech activity of a modern native speaker. The diversity of the content and volume of linguistic corpora makes it possible to obtain accurate data on the actual use of certain terms. The corpus becomes an important tool for analyzing the functioning of these terms both in the field of linguistics (morphology, syntax, vocabulary)

and in the theory and practice of translation, allowing us to determine the style of their formal or informal use.

Keywords: corpus linguistics, language corpus, methodology, representativeness, classification, criteria

Введение. История развития любого национального языка связана с меняющимися взглядами на его природу, сущность, структуру, общественную значимость и функции в глобальном контексте. Эта ситуация вполне оправдана, так как для эволюции научной лингвистической парадигмы характерна смена точек зрения, подходов и аспектов исследования. Лингвистика XIX века изучала язык как самостоятельный объект, в то время как в XX и XXI веках акцент сместился на практическое применение теоретических знаний. В современном языковедении текст и дискурс как объекты исследования определили сложность поиска научного материала, что повлекло за собой необходимость оптимизации работы с языковым контентом. В результате на пересечении лингвистики и программирования появилось новое направление — корпусная лингвистика, которая становится все более востребованной среди современных языковедов. С учетом быстрого развития информационных технологий, сбор и анализ практических данных и разнообразных текстов на разных языках приобретают новое значение и требуют тщательной разработки принципов и механизмов, основанных на основных положениях корпусной лингвистики. Профессор Г.П. Мельников считает, что любое исследование, проводимое лингвистом, должно быть ориентировано хотя бы на следующие этапы работы:

- 1) выбор принципов и критериев для классификации изучаемых объектов;
- 2) процесс распределения объектов по категориям согласно указанным критериям;

3) анализ, интерпретация и объяснение результатов распределения объектов по категориям, а также причин такого распределения.

Академик РАН В.А. Плунгян, руководитель кафедры теоретической и прикладной лингвистики МГУ им. М.В. Ломоносова, описывает корпусную лингвистику как «быстроразвивающееся» и «сверхсовременное» направление. Корпусная лингвистика обладает значительным исследовательским потенциалом, однако исследования российских и зарубежных ученых показывают различия в подходах к созданию и использованию корпусов в России и за рубежом.

На сегодня корпусную лингвистику можно считать одним из ключевых ресурсов для изучения языка и его описания. Что касается компьютерной лингвистики, то создание корпуса является основой этой дисциплины, позволяющей разрабатывать автоматизированные приложения для обработки текстов и других языковых проявлений.

Использование корпуса для изучения языковых явлений рассматривается как эмпирическая методология, основанная на фактических данных и образцах языкового общения. В этом контексте корпус представляет собой совокупность данных в самом широком смысле этого термина. Например, Словарь Испанской Королевской Академии, главного института в области языков Испании, описывает корпус как «наиболее обширный и упорядоченный набор данных или коллекцию технических, литературных и других текстов, которые могут стать основой для исследования». «Толковый словарь русского языка» определяет корпус как полное собрание определенных текстов.

Внедрение компьютерных технологий в процесс сбора, упорядочивания и обработки данных стало важным фактором, который современно трансформировал задачу создания корпусов, превратив отдельные эксперименты в целостную научную методологию и дисциплину, обозначаемую как корпусная лингвистика.

Некоторые ключевые события, заложившие основу корпусной лингвистики и способствовавшие ее развитию как научного направления, можно выделить. Так, испанская исследовательница М. Вильяндре Льямарес указывает, что до XIX века корпус в лингвистике понимался как:

- а) совокупность письменных текстов;
- б) объект изучения в контексте мертвых языков (например, латыни или санскриту);
- в) объект, к которому лингвисты могли обратиться только через методы корпусного анализа, так как не имели возможности собирать языковые данные от живых носителей.

В первой половине XX века американская структурная лингвистика заложила основы корпусной лингвистики как эмпирической методологии, основанной на наблюдении языковых данных. Однако термин «корпусная лингвистика» появился значительно позже, в начале 80-х годов XX века. Исследователи того времени считали, что корпус является единственным инструментом, подходящим для изучения языков, и утверждали, что он способен предоставить все необходимые данные для полной дескрипции конкретного языка. Эта новая концепция корпуса, известная как «структурный корпус», имела следующие характеристики:

- 1) совокупность устных образцов или письменных транскрипций (данные);
- 2) определением цели стало изучение живых языков, ранее не задокументированных в письменной форме (таких как языки американских индейцев);
- 3) необходимость сборов, так как получение устных образцов было единственным способом «доступа» к этим языкам;

4) сосредоточение на фонетических и (морфо)фонологических аспектах как уровнях, позволяющих провести «инвентаризацию» всех элементов, учитывая их законченную природу;

5) игнорирование репрезентативности полученных результатов. Поскольку анализ данных проводился вручную и визуально, не было возможности работать с большим объемом информации, что стало основанием для критики данной методологии и ее признания необъективной.

6) Многие лингвисты, включая Н. Хомского, придерживавшегося рационалистского подхода в языковых исследованиях, способствовали тому, что эмпирическая методология корпуса американских структуралистов в 1960—70-е годы уступила место другому подходу: рационализму, основанному на «интуиции лингвиста», как утверждал Н. Хомский. Он критиковал корпусную лингвистику с теоретической точки зрения, считая, что единственным ресурсом лингвиста являются его интуитивные ощущения, представляющие собой главный критерий в исследованиях; синтаксис, по его мнению, должен быть основным объектом изучения.

Основываясь на необходимости репрезентативности корпуса, Д. Байбер отмечает, что если «общий язык» является абстрактной категорией, а язык функционирует как система различных жанров и/или стилей, то референтный корпус должен включать все стили и жанры речи, а также региональные диалекты. В контексте социальной представленности языка он утверждает, что корпуса должны зафиксировать диалекты, социолекты и профессиональные языки, или языки для специальных целей. Д. Байбер уточняет, что язык должен рассматриваться в историческом аспекте, то есть включать тексты всех известных исторических эпох. Таким образом, он считает, что репрезентативность корпуса связана со сбалансированностью и пропорциональной представленностью жанров и стилей

языка, отражающих все слои общества, соответствующие реальной ситуации. В целом репрезентативность, по мнению Д. Байбера, заключается в включении в корпус текстов широкого спектра функциональных стилей и жанров.

П. Бейкер добавляет, что понятие репрезентативности связано с валидностью, то есть с соответствием полученных данных реальному состоянию языка в данной области использования. Исследователи считают, что полная репрезентативность в корпусах недостижима. Е.А. Красина также замечает, что словесные образы и сложные образные структуры управляются глубинными ассоциативными связями, которые проявляются на различных уровнях динамической структуры художественного текста.

Корпусная лингвистика как отдельная дисциплина окончательно сформировалась в 90-х годах XX века. В этот период электронные корпуса стали необходимым инструментом для изучения языка, разработки лингвистических гипотез и построения систем обработки естественного языка. На «возрождение» корпусной лингвистики, по нашему мнению, оказали значительное влияние многие ученые, среди которых стоит выделить Дж. Лича, начавшего дискуссию с критикой теоретических и практических аспектов идей Н. Хомского и Д. Аберкромби. Если в 1960-х годах эта критика была отчасти обоснованной, то в наши дни, согласно Дж. Личу, благодаря развитию компьютерных технологий, для создания корпусов выделяются несколько основных аргументов.

1) Корпус выступает в качестве научной методологии и поэтому имеет неоспоримое преимущество перед интуицией, так как результаты могут быть проверяемыми и не зависят от примеров, созданных лингвистами, заинтересованными в определенных результатах. Кроме того, в сфере количественных данных, таких как частотность, интуитивные ощущения являются неприемлемыми, так как наша восприятие частоты полностью субъективно.

2) Грамматическая структура текстов в корпусе позволяет ему отражать языковую компетенцию. Н. Хомский утверждал, что корпуса, как образцы речевых актов и языкового использования, не могут служить индикатором языковой компетенции. Тем не менее, исследования В. Лабова доказали высокий уровень грамматических последовательностей в корпусах.

3) Важность количественных данных: корпуса представляют собой уникальный источник для их получения.

4) Если структура корпуса научно обоснована, то данные о частотности употребления будут представлять собой достоверное отражение всего языка в целом.

5) Использование компьютеров опровергает мнение о применении «псевдонаучных» методов.

6) Компьютерная обработка больших объемов информации с низкими затратами и высокой скоростью позволяет минимизировать вероятность субъективных ошибок, связанных с человеческим фактором.

Термин «корпусная лингвистика» прочно входит в научный обиход после 1984 г., когда Дж. Аартс и В. Мейхс опубликовали работу ‘Corpus Linguistics In Recent Developments in the Use of Computer Corpora’ [24]. С этого момента термин начинает использоваться в своем современном значении. На наш взгляд, становлению этого научного направления способствовали следующие факторы: 1) расцвет прикладной лингвистики в целом и компьютерной лингвистики в частности, что сделало очевидным необходимость сбора и изучения данных

Использования языковых средств в речевой деятельности как носителями, так и не носителями языка. Эта необходимость объясняется тем, что, с одной стороны, корпуса отражают вариативность языка, с другой — могут фиксировать новые структуры или те конструкции, которые не соответствуют теоретическим

дескрипциям. К тому же в ситуациях с не носителями языка корпусы являются подлинным образцом возможных применений языка в речи; 2) эклектичность и неоднозначность понятия и его применения: использование корпуса в современном понимании понятия не противоречит аналитическому мнению лингвиста; сам по себе ни корпус (позиция американских структуралистов), ни интуиция идеального говорящего (слушателя) (по Н. Хомскому) не являются самодостаточными для объяснения лингвистических феноменов. В настоящее время признан тот факт, что изучение корпуса как набора текстов невозможно без интуиции и способности к интерпретации ученого аналитика, который использует свои знания языка (как носитель или компетентный не носитель языка), а также без владения им знаниями относительно языковой структуры (как лингвиста); 3) значительная доступность электронных корпусов благодаря Интернету; 4) развитие новых технологий информатизации текстов, таких, как оптическое распознавание знаков, автоматический диктант и т.д.; 5) значимость количественных данных в изучении определенных языковых аспектов; 6) необходимость создания более обширных глоссариев и словарей для поддержки компьютерных систем, которые способны работать с текстом любого рода, с субъязыками, жаргонами и разновидностями языка для специальных целей (например, медицинские или юридические тексты)

Литература.

1. Мельников Г.П. Системная типология языков: Принципы, методы, модели. М.: Наука, 2003.
2. Плунгян В.А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. 2008. № 2(16). С. 7—20.
3. Moure T., Llisterri, J. Lenguaje y nuevas tecnologías: el campo de la lingüística computacional” // M. Fernández Pérez (coord.) Avances en Lingüística aplicada. Santiago

de Compostela: Universidade de Santiago de Compostela, Servicio de Publicacións e Intercambio Científico, 1996. P. 147—227.

4. Real Academia Española. Diccionario de la lengua española. Madrid: Espasa, 2001. Режим доступа: <https://dle.rae.es/corpus> (дата обращения: 10.11.2020).

5. Ушаков Д.Н. Толковый словарь русского языка. Режим доступа: <https://gufo.me/dict/usakov/corpus> (дата обращения: 29.10.2020).

6. Villayandre Llamazares M. Lingüística con corpus // Estudios humanísticos. Filología. 2008. № 30. P. 329—349.

7. McEner T. Corpus Linguistics // The Oxford Handbook of Computational Linguistics / R. Mitkov (ed.). Oxford: Oxford University Press, 2003. P. 448—463.

8. McEner T., Wilson, A. Corpus Linguistics. Edinburg: Edinburgh University Press, 1996.

9. McEner T., Wilson A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 2001.

10. McEner T., Xiao R., Tono Y. Corpus-Based Language Studies. An advanced resource book. London—New York: Routledge, 2006. Режим доступа: <https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/chapters/B03.pdf> (дата обращения: 29.10.2020).

11. Chomsky N. Quine's empirical assumptions // Words and objections. Essay on the Work of W.V Quine / D. Davidson & J. Hintikka (eds.). Dordrecht: D. Reidel, 1969. P. 53—68.

12. Chomsky N. Language and mind. Cambridge, 2006.