O‘ZBEKISTON MILLIY UNIVERSITETI XABARLARI, 2025, [1/12/1]
ISSN 2181-7324

FILOLOGIYA
http://journals.nuu.uz
Social sciences

УДК:81’33:81’37

Интизор ДЖУМАНИЯЗОВА,
Преподаватель университета Маъмуна
E-mail: intizor1828@gmail.com

На основе рецензии к.ф.н. PhD, У.Р.Махмудова

ЗНАЧЕНИЕ И ФУНКЦИОНАЛЬНЫЙ ПОТЕНЦИАЛ АВТОРСКИХ КОРПУСОВ

Аннотация

Статья посвящена значению и функциональному потенциалу авторских корпусов текстов в современном научном, образовательном и культурном пространстве. Рассматривается роль цифровых технологий в создании и использовании авторских корпусов, подчеркивается их значение для корпусной лингвистики, педагогики, филологии и гуманитарной науки в целом. Автор анализирует отечественный и международный опыт создания авторских корпусов, включая примеры из Национального корпуса русского языка, а также работы по А. П. Чехову, Ф. М. Достоевскому. Особое внимание уделяется лексико-грамматическим и стилистическим характеристикам текстов, методам их разметки, а также проблемам доступности, стандартизации и правового регулирования.

Ключевые слова: Авторские корпуса, лексический анализ, лингвистические исследования, лемматизация, токенизация, корпусный словарь.

MUALLIFLIK KORPUSINING AHAMIYATI VA FUNKSIONAL IMKONIYATLARI

Annotatsiya

Ushbu maqola zamonaviy ilmiy, ta’lim va madaniy kontekstlarda muallif tomonidan yaratilgan korpusning ahamiyati va funksional imkoniyatlarini o‘rganadi. Muallif tomonidan yaratilgan korpuslarni yaratish va ulardan foydalanishda raqamli texnologiyalarning roli ko‘rib chiqilib, ularning korpus lingvistika, pedagogika, filologiya va umuman gumanitar fanlar uchun ahamiyati ko‘rsatilgan. Mualliflik korpusini shakllantirishda mahalliy va global tajribalar, xususan, "Rus tili milliy korpusi" materiallari, shuningdek, A.P.Chexov va F.M. Dostoevskiy asarlari namunalari o‘rganiladi. Matnlarning lug‘aviy, grammatik va uslubiy jihatlari, ularni tahlil qilish usullari, qo‘llanilishi, standartlashtirish va huquqiy asoslari masalalariga alohida urg‘u beriladi.

Kalit so‘zlar: Muallif korpusi, leksik tahlil, lingvistik tadqiqot, lemmatizatsiya, tokenizatsiya, korpus lug‘ati.

THE SIGNIFICANCE AND FUNCTIONAL POTENTIAL OF AUTHOR'S CORPSES

Annotation

This article examines the significance and functional potential of author-created corpora in contemporary scientific, educational, and cultural contexts. The role of digital technologies in the creation and use of author-created corpora is examined, highlighting their importance for corpus linguistics, pedagogy, philology, and the humanities in general. The author analyzes domestic and international experience in creating author corpora, including examples from the National Corpus of the Russian Language, as well as works on A.P. Chekhov and F.M. Dostoevsky. Particular attention is paid to the lexical, grammatical, and stylistic characteristics of texts, methods of their annotation, as well as issues of accessibility, standardization, and legal regulation.

Key words: Author corpora, lexical analysis, linguistic research, lemmatization, tokenization, corpus dictionary.

Уникальные авторские коллекции текстов обладают способностью всецело и безошибочно репрезентировать языковой стиль писателя, что выгодно отличает их от прочих цифровых баз данных. Во многих уголках планеты ценность таких корпусов возрастает ввиду их обширного и всестороннего информационного потенциала. Нельзя не подчеркнуть роль цифровых инструментов в их создании и использовании.

В наш век информации сложно представить учебный процесс без использования лингвистических корпусов. Подобно точным наукам, лингвистика осуществляет масштабные и глубокие исследования, опираясь на конкретные данные и эмпирические подтверждения. Лингвистические корпуса предоставляют необходимые ресурсы для такой работы, обеспечивая надежную основу для анализа и исследований.

Благодаря корпусной лингвистике мы получаем возможность детально исследовать лексический состав языка. Также упрощается отслеживание языковых изменений, включая выявление устаревших, вышедших из употребления слов, а также новых и заимствованных лексических единиц.

В российской лингвистике существует значительное количество специализированных работ, в которых анализируются языковые характеристики произведений конкретных авторов. Отдельные исследования посвящены лингвистическим особенностям текстов, написанных с использованием рунического и уйгурского письма. Международный опыт показывает, что корпусные исследования особенно важны и эффективны в этих областях. Они предоставляют обширный материал и полезны для анализа и интерпретации текстов.

Применение программного обеспечения для компьютеров обеспечивает получение достоверных сведений. Одной из ключевых причин повышенного внимания к исследованиям корпусов текстов является их востребованность в качестве удобного и распространенного инструмента обучения и работы в сфере образования. Рост интереса обусловлен двумя факторами: первоначальным увеличением спроса и

последующей разработкой компьютерной техники, способной обрабатывать значительные объемы текстовых данных. Это состояние считается типичным и ожидаемым.

Часть информации, искомой в книгах, неизвестна читателю, что создает определенные трудности с точки зрения времени и получения точной информации при изучении этого материала.

Россия занимает передовые позиции в области разработки авторских текстовых коллекций. Творчество Антона Павловича Чехова, признанное во всем мире и переведенное на множество языков, сохраняет свою актуальность и влияние на современную культуру. Это обусловило повышенный интерес к созданию сначала авторских словарей, а затем и полнотекстовых корпусов, посвященных его произведениям. Все работы Чехова представлены в Национальном корпусе русского языка[6]. Из-за сложностей с доступностью тексты были отсканированы и собраны из различных онлайн-источников. Важно понимать, что ни один словарь не способен полностью отразить все многообразие лексики языка. Поэтому пользователям предлагается использовать корпус произведений Чехова для изучения контекстов употребления слов и выражений, а также для развития навыков поиска и анализа. Интерес исследователей к этой области постоянно растет. В настоящее время корпус Чехова включает в себя все прозаические произведения, пьесы, а также публицистические работы писателя. Переписки, дневниковые записи и материалы из "Острова Сахалин" и "Полное собрание сочинений"[11] пока не включены в состав корпуса.

Творческое наследие Антона Павловича Чехова поражает своим объемом: он создавал как короткие рассказы, так и масштабные романы, пьесы для театра, а также занимался публицистикой и писал критические статьи. Корпусный словарь Чехова содержит около 36 153 лемм[6] или лексем, что соответствует 1 381 000 словарных знаков (120 000 словоформ). Лексические единицы используются в 168 000 предложений, поэтому средняя длина предложения составляет около 8 слов (8,22 слова), не считая числительных. Следует отметить, что включение необработанных материалов может существенно расширить этот показатель. При сравнении произведений Чехова со словарным запасом Гёте выяснилось, что словарный запас Гёте составлял около 90 000 лексем, несмотря на то, что творчество Гёте охватывало почти 50 лет. К сожалению, из-за нескольких реформ орфографии и работы редакторов на протяжении многих лет язык текстов корпуса не в полной мере соответствует языку автора, поскольку после перевода текстов в электронный вид была проведена большая техническая работа, включая токенизацию и лемматизацию. Лексемы авторского корпуса Чехова представлены техническими, современными и устаревшими словами, иноязычными словами, а также авторскими неологизмами,[5] введенными в общий оборот самим автором. База данных содержит 4840 названий с нелемматизированными иноязычными словоформами и частично лемматизированными словами в текстах на латинской графике. В сборнике представлены прозаические произведения, драматургия и публицистика. Корпус произведений А. П. Чехова предоставляет прекрасную возможность для изучения его лексики, помогает лучше понять язык и мировоззрение писателя, а также исторический период конца XIX – начала XX веков.

Позднее в результате ряда исследований были созданы авторские электронные словари и авторские корпуса, отражающие культуру, обычаи и историю русского народа. Например, в Московском государственном университете имени М.В. Ломоносова подготовлены такие словари, как «Словарь хрустального языка», «Словарь языка комедии Грибоедова», «Язык комедии Гоголя «Ревизор»», «Язык комедии Фонтского». Этот эксперимент[8] весьма популярен в русской лексикографии. Словари «Словарь сочинений Ломоносова»[1] и «Пан Тадеуш» А. Мицкевича также были созданы как последовательное продолжение вышеуказанных исследований. Среди исследований в этой области большое значение имеют работы О. В. Кукушкиной и А. Поликарпова. Анализ корпусов авторских текстов показывает, что они лингвистически комплементарны. Корпуса авторских текстов Ф. Достоевского и А. Чехова имеют как общие, так и специфические черты. Существующие электронные издания произведений Ф. Достоевского включают электронную коллекцию. Необходимость создания лингвистической базы данных обусловлена тем, что язык Достоевского до сих пор не изучен детально и в полной мере. Лексические, грамматические, морфологические, морфемные, вербальные и синтаксические особенности произведений писателя не были систематизированы, а общее употребление слов, словосочетаний и лексики в различных жанрах не изучалось как единая система. Становится очевидным, что эта задача может быть решена путём создания корпуса произведений Ф. Достоевского. Корпус текстов Ф.М. Достоевского был создан в качестве источника для составления словаря языка Достоевского, а параметры составления корпуса основывались на структуре и содержании словарной статьи. Для полного охвата лексики Ф.М. Достоевского в корпус включены все литературные и публицистические произведения автора. Сегодня тексты писателя рассматриваются как его литературные произведения, публицистика и эпистолярное наследие на различных электронных носителях. По авторству Достоевского создано несколько словарей, в том числе частотный словарь,[2] составленный А. Я. Шайкевичем. Корпус также включает идиоматическую базу данных, на основе которой составлен словарь авторских идиом. Корпус Ф.М. Достоевского также представлен на CD-ROM и он называется: «Достоевский: тексты, исследования, материалы».

В современном обществе, где информация и язык играют ключевую роль в формировании культурной, научной и образовательной среды, авторские корпуса приобретают особое значение как важный лингвистический, гуманитарный и социокультурный ресурс. Они представляют собой не только хранилище текстов конкретного автора, но и многофункциональный инструмент, позволяющий исследовать динамику языка, индивидуальный стиль, культурные ценности и идеологические ориентиры эпохи. Значение авторских корпусов выходит далеко за рамки лингвистических исследований. Их использование находит применение в самых разных сферах общественной жизни. Таким образом, авторские корпуса становятся не просто научным продуктом, но и общественно значимым инструментом, влияющим на формирование языковой культуры, образовательных стандартов, сохранение культурной памяти и развитие технологий. Их грамотное построение и осознанное использование способны внести существенный вклад в интеллектуальное развитие как отдельных пользователей, так и общества в целом.

ЛИТЕРАТУРА

1. Волков С.С., Матвеев Е.М. “О проекте словаря “Риторика” М.В.Ломоносова. Доклад (ILI RAN). (Elektron resurs: www.ruslang.ru/seminar_aut_lexocorg 250214) 159 Гик А.В. “О работе “Конкордансом к стихотворениям М. Кузмина”: том четвёртый” (Elektron resurs: www.ruslang.ru/seminar_aut_lexocorg)
2. Заботкиной В. И “Методы когнитивного анализа семантики слова компьютерно-корпусный подход” / Под. общ. ред. – Москва: Языки славянской культуры, 2015. – 344 с.
3. Захаров В. П. Корпусная лингвистика. – СПб. Санкт-Петербургский ГУ, 2005. – С. 6.
4. Захаров В., Богданова С. Корпусная лингвистика. – СПб. Санкт-Петербургский ГУ, 2020. – С. 59.
5. Кукушкина О. В., Рюдигер Д.Ю., Суровцева Е.В., Лапонина В.V под.ред, проф. Поликарпова А.А (2012), Частотный грамматико-семантический словарь языка художественных произведений А.П.Чехова (с электронным приложением) М.: МАКС Пресс.
6. Потемкин С. Б. “Авторский корпус и словарь языка Антона Чехова (Электрон ресурс://https: istina.msu.ru.
7. Поликарпов А.А. Корпусная информационно исследовательская система. // Электронная энциклопедия языка: Вып. 1. Стихи и драмы А.С. Пушкина. Путеводитель по Пушкину. – М, 2006. https://lex.philol. msu. ru/ proekty/kiisa/
8. Сироткина Т. А. Национальный корпус русского языка как материал для создания этнонимического словаря. // Национальный корпус русского языка и проблемы гуманитарного образования. – М.: Наука, 2007. – С. 230-234.
9. Словарь языка Достоевского. Идиоглоссарий. А-В/ Российская академия наук институт русского языка им В. В. Виноградова; гл. редактор чл. корр. Караулов М. 2008.
10. Хамроева Ш. Общее и уникальное в корпусе авторских текстов // Международный журнал речевого искусства, 2020. Том 3, Выпуск 2, – С. 86.
11. Чехов А. П. Полное собрание сочинения и писем в 30 томах. – М.: Наука, 1977.
12. Foydalanilgan adabiyotlar ro‘yxati:
13. Volkov S.S., Matveev E.M. "M.V. Lomonosovning "Ritorika" lug'ati loyihasi to'g'risida. Hisobot (ILI RAS). (Elektron resurs: www.ruslang.ru/seminar_aut_lexocorg 250214) 159 Gik A.V. To'rt" (Elektron resurs: www.ruslang.ru/seminar_aut_lexocorg)
14. Zabotkina V.I. "So'z semantikasini kognitiv tahlil qilish usullari: kompyuter-korpus yondashuvi" / Ed. – Moskva: Slavyan madaniyati tillari, 2015. – 344 b.
15. Zaxarov V.P. Korpus tilshunosligi. – Sankt-Peterburg davlat universiteti, 2005. – B. 6.
16. Zaxarov V., Bogdanova S. Korpus tilshunosligi. - SPb. Sankt-Peterburg davlat universiteti, 2020. - B. 59.
17. Kukushkina O. V., Rüdiger D. Yu., Surovtseva E. V., Laponina V. V. ed., prof. Polikarpova A. A (2012), A. P. Chexov badiiy adabiyoti tilining chastotali grammatik-semantik lug'ati (elektron qo'shimcha bilan) M.: MAKS Press.
18. Potemkin S. B. “Anton Chexov tilining mualliflik korpusi va lug'ati (Elektron manba: https://https: istina.msu.ru.
19. Polikarpov A.A. Korpus axborot tadqiqot tizimi. // Elektron til ensiklopediyasi: 1-son. A.S. Pushkinning she’r va dramalari. Pushkin bo’yicha qo'llanma. – M, 2006. https://lex.philol. msu. ru/proekty/kiisa/
20. Sirotkina T.A. Rus tilining milliy korpusi etnonimik lug'at yaratish uchun material sifatida. // Rus tilining milliy korpusi va gumanitar ta'lim muammolari. – M.: Nauka, 2007. – B. 230-234.
21. Dostoevskiy tilining lug'ati. Idioglossarium. A-B/ Rossiya Fanlar akademiyasi, V.V. Vinogradov nomidagi rus tili instituti; Bosh muharrir muxbir a’zo Karaulov M. 2008 yil.
22. Hamroeva Sh. Mualliflik korpusining mushtarak va o’ziga hos jihatlari // So’z san’ati xalqaro jurnali, 2020. 3-jild, 2-son, – B. 86.
23. Chexov A. P. 30 jildlik asarlar va maktublarning to‘liq to‘plami. - M.: Nauka, 1977 yil.
24. List of used literature:
25. Volkov S.S., Matveev E.M. “On the Project of the Dictionary “Rhetoric” by M.V. Lomonosov. Report (ILI RAS). (Elektron resurs: www.ruslang.ru/seminar_aut_lexocorg 250214) 159 Gik A.V. “On the Work “Concordance to the Poems of M. Kuzmin”: Volume Four” (Elektron resurs: www.ruslang.ru/seminar_aut_lexocorg)
26. Zabotkina V.I. “Methods of Cognitive Analysis of Word Semantics: A Computer-Corpus Approach” / Ed. – Moscow: Languages of Slavic Culture, 2015. – 344 p.
27. Zakharov V.P. Corpus Linguistics. – St. Petersburg State University, 2005. – P. 6.
28. Zakharov V., Bogdanova S. Corpus linguistics. - SPb. St. Petersburg State University, 2020. - P. 59.
29. Kukushkina O. V., Rüdiger D. Yu., Surovtseva E. V., Laponina V. V. ed., prof. Polikarpova A. A (2012), Frequency grammatical-semantic dictionary of the language of A. P. Chekhov's fiction (with electronic supplement) M.: MAKS Press.
30. Potemkin S. B. “Author's corpus and dictionary of the language of Anton Chekhov (Electronic resource: https://https: istina.msu.ru.
31. Polikarpov A.A. Corpus information research system. // Electronic encyclopedia of language: Issue 1. Poems and dramas by A.S. Pushkin. Guide to Pushkin. – M, 2006. https://lex.philol. msu. ru/proekty/kiisa/
32. Sirotkina T.A. National corpus of the Russian language as material for creating an ethnonymic dictionary. // National corpus of the Russian language and problems of humanitarian education. – M.: Nauka, 2007. – Pp. 230-234.
33. Dictionary of the language of Dostoevsky. Idioglossarium. A-B/ Russian Academy of Sciences, V.V. Vinogradov Russian Language Institute; Editor-in-Chief Corresponding Member Karaulov M. 2008.
34. Khamroeva Sh. The General and the Unique in the Corpus of Author's Texts // International Journal of Speech Art, 2020. Volume 3, Issue 2, – P. 86.
35. Chekhov A. P. Complete Works and Letters in 30 Volumes. – M.: Nauka, 1977.