Научная статья
УДК 81'1
ББК 81.11
О 74

DOI: 10.53598/2410-3489-2023-1-312-118-124

О перспективах корпусных исследований и практике их применения в лингвистике

(Рецензирована)

Георгий Анатольевич Осипов
Адыгейский государственный университет, Майкоп, Россия,
osipoff.georgy@yandex.ru

Аннотация:
Анализируются магистральные направления современных корпусных исследований и возможности их применения в современной лингвистике и филологии. Описываются основополагающие положения корпусной лингвистики в рамках отечественного и зарубежного языкознания. Проводится классификация лингвистических корпусов различных типов и обзор их инструментария. Рассматриваются проблемы анализа массива текстовых данных и индивидуальных текстовых корпусов. Проводится статистический анализ объема публикаций в наукометрических базах данных. Делаются выводы о перспективах корпусных методик в современных языковых исследованиях.

Ключевые слова: корпус, корпусная лингвистика, лингвистический корпус, корпусное исследование, корпус текстов, массив текстовых данных, автоматизированная обработка данных.

Для цитирования: Осипов Г.А. О перспективах корпусных исследований и практике их применения в лингвистике // Вестник Адыгейского государственного университета. Сер.: Филология и искусствоведение, 2023. Вып. 1 (312). С. 118-124. DOI: 10.53598/2410-3489-2023-1-312-118-124.

Original Research Paper

On the prospects of corpus research and the practice of its application in linguistics

Georgy A. Osipov
Adyghe State University, Maykop, Russia,
osipoff.georgy@yandex.ru

Abstract:
The study aims to analyze main fields of modern corpus research and their application in modern linguistics and philology. The fundamentals of corpus linguistics in the framework of domestic and foreign linguistics are described. The study provides classification of linguistic corpora of various types and a review of their tools. The problems of analyzing an array of text data and individual text corpora are considered. The study carries out statistical analysis of the volume of publications in scientometric databases. The prospects of corpus methods in modern language studies have been summarized.

Keywords: corpus, corpus linguistics, linguistic corpus, corpus research, corpus of texts, array of text data, automated data processing.

For citation: Osipov G.A. On the prospects of corpus research and the practice of its application in linguistics//Bulletin of Adyghe State University, Ser.: Philology and Art Criticisms, 2023. No.1 (312). P. 118-124. DOI: 10.53598/2410-3489-2023-1-312-118-124.

Введение.

Зарождение интереса к исследованию больших объемов текстовой информации относят к середине XX века, когда впервые в истории начали появляться технические средства и возможности для создания автоматизированных систем обработки текстовых данных. «Рост объема информации в современном информационном обществе затрудняет обработку, хранение, распространение, поиск информации, то есть результативную работу с ней. Данную проблему можно решить с помощью перевода информации в электронную форму и использования информационно-коммуникационных технологий» [1: 133]. Развитие компьютерных технологий и возрастание вычислительных мощностей дало толчок в развитии направления языковых исследований, получившему название корпусная лингвистика [2]. Термин введён в широкое употребление в 1960-е годы, что явилось результатом создания массивов текстовых данных, содержащих многочисленные примеры актуальных языковых контекстов. В основе «фундаментального множества информационных технологий лежит коммуникация, неразрывно связанная с языком и, следовательно, лингвистикой» [3: 102]. Считается, что первым языковым корпусом в прямом смысле этого слова стал созданный в 1961 г. в университете Брауна (США) корпус английского языка, содержавший 500 фрагментов текстов по 2 тысячи слов в каждом. Позднее общепризнанным международным стандартом представительности стал объем языкового корпуса равный 1 млн. словоупотреблений. В.В. Рыков дает следующее определение корпуса: «некоторое собрание текстов, в основе которого лежит логический замысел, логическая идея, объединяющая эти тексты и воплощенная в правилах организации текстов в корпус, алгоритме и программе анализа корпуса текстов, сопряженной с этим идеологии и методологии» [4].

Языковой корпус, созданный в университете Брауна, стал моделью и прообразом других национальных корпусов, создание которых продолжается по сей день. Актуальность исследования заключается в том, что современные технические средства дают возможность создания индивидуально-авторских текстовых корпусов, а также автоматизированной системы для их обработки, позволяющей мгновенно получать разнообразные данные о тексте и его содержании, а также представлять их в удобном для дальнейшего использования виде.

Материалы и методы.

В исследовании использовались такие методы, как лингвистическое описание, дефиниционный метод, метод систематизации и классификации материала, наблюдение, а также, метод статистической обработки данных и квантитативный сравнительный анализ.

В качестве материала исследования были использованы данные из наукометрических баз данных, таких как Elibrary (РИНЦ и Science index), WoS, Scopus, Google Scholar, а также данные национальных языковых корпусов, таких как BNC, COCA, Национальный корпус русского языка.

Результаты.

Одним из самых популярных национальных корпусов является BNC (British National Corpus) [5]. Данный корпус содержит наиболее

широко используемые онлайн-базы данных, которые применяются для самых разных целей преподавателями и исследователями в университетах по всему миру [6]. Кроме того, данные корпуса (например, полнотекстовые, частотность слов) использовались широким кругом компаний в самых разных областях, особенно в области технологий и изучения языков.

Британский национальный корпус (BNC) был первоначально создан издательством Oxford University press в 1980-х – начале 1990-х годов и содержит 100 миллионов слов текста из широкого спектра жанров (например, разговорный, художественная литература, журналы, газеты и академические издания).

В состав данного корпуса входят следующие разделы (см. табл. 1):

Таблица 1.

Разделы корпуса BNC с количеством слов

Название раздела	Количество слов	Период охвата
News on the Web (NOW)	17,3 млрд.	2010-2023
iWeb: The Intelligent Web-based Corpus	14 млрд.	2017
Global Web-Based English (GloWbE)	1,9 млрд.	2012-2013
Wikipedia Corpus	1,9 млрд.	2014
Corpus of Contemporary American English (COCA)	1 млрд.	1990-2019
Corpus of Historical American English (COHA)	475 млн.	1820-2019
The TV Corpus	325 млн.	1950-2018
The Movie Corpus	200 млн.	1930-2018
Corpus of American Soap Operas	100 млн.	2001-2012
Early English Books Online	755 млн.	1470-1690
TIME Magazine Corpus	100 млн.	1923-2006

Еще одним примером весьма представительного корпуса можно считать корпус современного американского английского языка (Corpus of Contemporary American English) [7]. Он является крупным и репрезентативным корпусом американского английского языка, а также наиболее широко используемым корпусом английского языка, и связан со многими другими корпусами английского языка. Эти корпуса ранее были известны как «Корпуса BYU», и они дают беспрецедентное представление о вариациях английского языка. Корпус содержит более миллиарда слов текста (более 25 миллионов слов в год с 1990 по 2019 год) из восьми жанров:

– устная речь;
– художественная литература;
– популярные журналы;
– газеты;
– академические тексты;

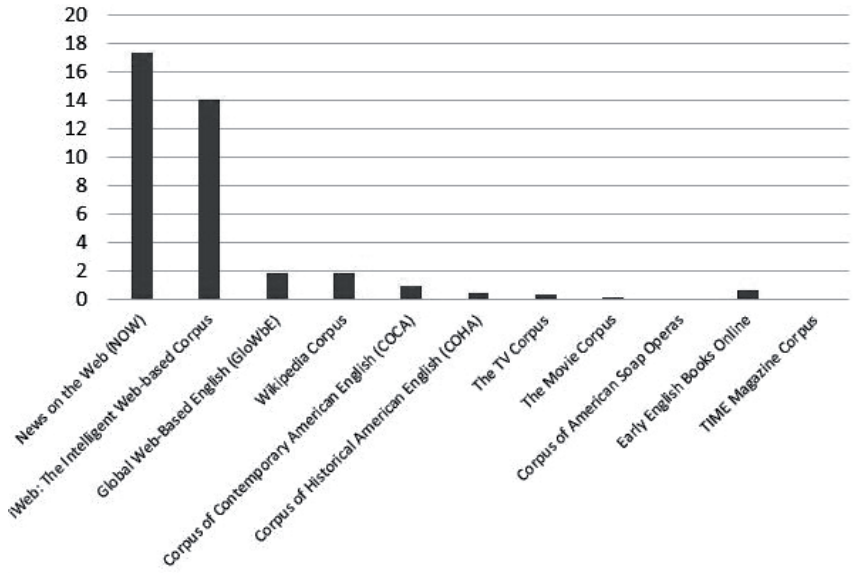

Диаграмма 1. Сравнение репрезентативности разделов BNC (млрд. слов).
Diagram 1. Comparison of the representativeness of BNC sections (billion words).

– субтитры к телевидению и фильмам;

– блоги;

– другие веб-страницы.

В качестве примера русскоязычного корпуса можно привести Национальный корпус русского языка [8], который позиционируется как представительная коллекция текстов на русском языке общим объемом более 2 млрд слов, оснащенная лингвистической разметкой и инструментами поиска. Он создан по образцу англоязычных корпусов и содержит следующие разделы (см. табл. 2):

Таблица 2.

Разделы Национального корпуса русского языка с количеством слов

Название раздела	Количество слов
Основной	374 млн.
Газетные	790 млн.
Синтаксический	1,5 млн.
Социальные сети	160 млн.
Устный	13 млн.
Акцентологический	133 млн.
Мультимедийный	5,5 млн.
Диалектный	599 тыс.
Поэтический	13 млн.
Исторические	14 млн.

В качестве доказательства востребованности данного направления в современной науке можно также привести количество публикаций, затрагивающих эти вопросы. Например, количество публикаций в наукометрической базе данных Elibrary (РИНЦ и Science index), в которых одним из ключевых слов является «языковой корпус», равняется 13078, с ключевым словом «корпус английского языка» – 11696, с ключевым словом «корпус русского языка» – 14925, с

ключевым словом «корпусная лингвистика» – 8573.

Обсуждение.

Обработка корпусных данных предполагает создание определенного алгоритма, которые даст решение ряду научных проблем как теоретического, так и прикладного характера в области математического лингвистического моделирования, построения лингвистической инфографики, а также автоматизированной обработки текстовых данных (разнородных текстовых массивов, языковых или индивидуально-авторских корпусов) по определенным критериям. Критерии анализа текста предполагают максимальную гибкость в зависимости от потребностей конкретного пользователя. В результате теоретического исследования, а также практической работы предполагается создание компьютерного алгоритма автоматизированного текстового анализа, позволяющего автору научной работы или рецензенту (эксперту / оппоненту) за короткий срок провести анализ текста любого объема по заданным параметрам.

Отличительной особенностью данного алгоритма является возможность добавления / изменения параметров анализа текста по запросу пользователя. Так, например, автору научной работы алгоритм даст возможность создать индивидуально-авторский корпус текстов (например, полнотекстовый корпус произведений какого-либо автора), проанализировать его по желаемым параметрам (графическим, лексическим, морфологическим и т.д.) и вывести результаты в виде графиков, таблиц, диаграмм или схем. Обработка и сравнение данных, полученных в результате анализа текста по всем возможным параметрам, даст возможность определить особенности такого сложного понятия, как идиостиль того или иного автора.

С другой стороны, алгоритм даст возможность специалисту, анализирующему готовую научную работу (научному руководителю, рецензенту, эксперту или оппоненту) практически мгновенно провести автоматизированную обработку текста по необходимым критериям и визуализировать полученные данные в виде процентных соотношений, коэффициентов, таблиц, диаграмм и т.д.

Создание подобного алгоритма предполагает решение следующих задач:

– определить возможности создания массива текстовых данных и персональных корпусов;

– определить методики обработки и анализа массива текстовых данных и персональных корпусов;

– определить возможности применения, принципы и критерии использования математических и статистических методов обработки и анализа массива текстовых данных;

– выработать методику применения средств математического моделирования лингвистических данных;

– выработать методику представления полученных данных в виде инфографики;

– определить потенциал, перспективы использования и сферы применения программного обеспечения и/или web-сервиса автоматизированной обработки текста.

Создание подобного алгоритма может иметь следующее практическое применение:

1. Возможность использования общеязыковых корпусов и создания персональных корпусов текста любого человека (автора, журналиста, политического деятеля, ученого и т.д.).

2. Независимость от языка, на котором написан текст, так как программное обеспечение дает возможность интеграции любых текстовых данных.

3. Возможность применения при написании научных работ так как позволяет анализировать массивы данных (к примеру, возможно проанализировать на тот или иной предмет исследования все произведения

автора сразу, вместо одного или нескольких). Это особенно актуально при исследовании концептов, где необходимо выявлять лексемы, номинирующие концепт или анализировать структуру концепта. Также подобный подход имеет практически безграничный потенциал для количественного анализа текста.

4. Возможность рецензирования и редактирования текстов любого рода по множеству критериев (к примеру, за счет мгновенного выявления в массиве текста жаргонизмов, сленгизмов, архаизмов, слов-паразитов, наукообразных терминов, сниженной или разговорной лексики, повторов и др.) Программное обеспечение должно позволять представлять полученные данные в виде схем, графиков, диаграмм, процентных соотношений и т.д.

5. Возможность выявления корреляция общеязыковых закономерностей и индивидуальных особенностей автора. Предполагается взять за основу общеязыковой корпус и провести статистическое сравнение по заданным критериям. Возможность на основании этих соотношений описать идиостиль автора (или любого человека).

6. Возможность выявления индивидуальных речевых особенностей (недостатков) любого человека и выработка рекомендаций по корректировке речи (к примеру, в сфере политики, журналистики и т.д.)

7. Перспективы практически безграничного совершенствования программного обеспечения за счет внедрения новых аспектов анализа текстовых данных.

Заключение.

Корпусная лингвистика как самостоятельный раздел прикладной лингвистики в настоящее время является одним из самых перспективных направлений развития языкознания. Электронные корпусы текстов получили огромное значение как в лингвистических исследованиях, так и в практике преподавания языков и переводоведения. Применение компьютерных технологий в области корпусной лингвистики открывает перед исследователем новые возможности. Объем текстов, анализируемых с помощью специально разработанных программ, позволяет получать статистические данные о языке, недоступные прежде.

Примечания:

1. Хурум Р.Ю. Использование информационно-коммуникационных технологий в лингвистике как фактор формирования информационных компетенции // Вестник Адыгейского государственного университета. Сер.: Филология и искусствоведение. Майкоп, 2017. Вып. 4. С. 133-138.

2. Баранов А.Н. Проблема репрезентативности корпуса данных (на примере политической метафорики) // Труды Международного семинара «Диалог-2001» по компьютерной лингвистике и ее приложениям. Аксаково, 2001.

3. Птущенко Е.Б. Перспективные информационные технологии как инструмент познания в лингвистике // Вестник Адыгейского государственного университета. Сер.: Филология и искусствоведение. Майкоп, 2017. Вып. 2. С. 102-108.

4. Рыков В.В. Корпус текстов как реализация объектно-ориентированной парадигмы // Труды международного семинара «Диалог-2002». URL: https://www.dialog-21.ru/digest/2002/articles/rykov/ (дата обращения: 15.02.2023).

5. The British National Corpus (BNC). URL: https://www.english-corpora.org/ (дата обращения: 15.02.2023).

6. McEnery T., Wilson A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 1999. 256 p.

7. The Corpus of Contemporary American English (COCA). URL: https://www.english-corpora.org/coca/ (дата обращения: 15.02.2023).

8. Национальный корпус русского языка. URL: https://web.ruscorpora.ru (дата обращения: 15.02.2023)

References:

1. Khurum R.Yu. The use of information and communication technologies in linguistics as a factor in the formation of information competences // Bulletin of the Adyghe State University. Ser.: Philology and the Arts. Maikop, 2017. Iss. 4. P. 133-138.

2. Baranov A.N. The problem of representativeness of the data corpus (based on political metaphors) // Proceedings of the International Seminar «Dialogue-2001» on Computational Linguistics and its Applications. Aksakovo, 2001.

3. Ptushchenko E.B. Perspective information technologies as a tool of knowledge in linguistics // Bulletin of the Adyghe State University. Ser.: Philology and the Arts. Maikop, 2017. Iss. 2. P. 102-108.

4. Rykov V.V. Corpus of texts as an implementation of the object-oriented paradigm // Proceedings of the international seminar «Dialogue-2002». URL: https://www.dialog-21.ru/digest/2002/articles/rykov/ (date of access: 15.02.2023).

5. The British National Corpus (BNC). URL: https://www.english-corpora.org/ (date of access: 15.02.2023).

6. McEnery T., Wilson A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 1999. 256 pp.

7. The Corpus of Contemporary American English (COCA). URL: https://www.english-corpora.org/coca/ (date of access: 15.02.2023).

8. National corpus of the Russian language. URL: https://web.ruscorpora.ru (date of access: 15.02.2023).

Статья поступила в редакцию 21.01.2023; одобрена после рецензирования 18.02.2023; принята к публикации 22. 03.2023.

The paper was submitted 21.01.2023; approved after reviewing 18.02.2023; accepted for publication 22.03.2023.