RUDN Journal of Language Studies, Semiotics and Semantics 2021 Vol. 12 No. 1 196—218
Вестник РУДН. Серия: ТЕОРИЯ ЯЗЫКА. СЕМИОТИКА. СЕМАНТИКА http://journals.rudn.ru/semiotics-semantics

DOI: 10.22363/2313-2299-2021-12-1-196-218
УДК 81

Научная статья / Research article

Корпусная лингвистика: теория vs методология

К.П. Чилингарян

Российский университет дружбы народов,
117198, Российская Федерация, Москва, ул. Миклухо-Маклая, 6
chilingaryan-kp@rudn.ru

Аннотация. Проведено комплексное исследование этапов становления и развития корпусной лингвистики. Целью статьи является анализ научных подходов к вопросу научной значимости рассматриваемой лингвистической дисциплины, а также выявление комплекса понятий и критериев, составляющих фундамент данного направления. Корпусная лингвистика представляет собой одну из наиболее перспективных и быстро развивающихся областей языковых исследований. Лингвистика XIX века ставила своей целью изучение языка как такового, лингвистика XXI века видит актуальность исследования не в выявлении абсолютных лингвистических категорий и значений, но в практическом применении лингвистических знаний. Актуальность представляемой статьи определяется тем, что в лингвистических корпусах заложен огромный потенциал, который еще не в полной мере осмыслен научным сообществом, хотя бы в силу того, что текст — основной объект корпусной лингвистики — в различных формах своей реализации представляет собой одну из главных составляющих системы языка и речемыслительной деятельности современного носителя языка. Содержание и объем лингвистических корпусов различного рода позволяет получать достоверную информацию об актуальном и реальном использовании того или иного термина: корпус становится инструментом анализа функционирования этого термина как в лингвистической области (морфологии, синтаксиса и лексики), так и в теории и практике перевода, идентифицируя регистр его формального или неформального узуса. Принципиальная новизна результатов данного исследования позволяет говорить о правомерности создания корпусных словарей и корпусных грамматик нового поколения, разработанных и верифицированных по отношению к конкретному фиксированному корпусу. Одновременно обосновывается положение о том, что корпусный характер словарей и грамматик повышает их надежность, достоверность и объективность и позволяет избежать субъективности, которая нередко свойственна исследованиям, опирающимся исключительно на интуицию лингвиста. Корпус является средой для получения новых научных данных, осмысление которых представляется приоритетным для современного лингвистического описания и абсолютно необходимым в научной деятельности современного исследователя. Новизна проведенного анализа заключается в том, подтверждена целесообразность корпусных исследований как сущностное требование времени, связанное с новым качеством лингвистической реальности и отвечающее потребностям современного общества. В статье рассматриваются основные этапы становления корпусной лингвистики как научного направления, характеризуются научные представления и подходы, присущие каждому из этих этапов, представляется обзор основных понятийных положений корпусной лингвистики в рамках отечественного и зарубежного языкознания. Автор

подробно анализирует полемику между представителями различных научных направлений и выявляет преимущества того или иного подхода, прослеживает сходства и различия между подходами к изучению корпусов на различных исторических этапах становления изучаемого научного направления. В фокусе обзора роль и место корпусных исследований языка в современной лингвистике, сопоставление аргументов pro и contra применения корпусных технологий в лингвистическом описании. Значительное внимание обращается на основные критерии классификации корпусов, предлагается краткий обзор наиболее известных в истории корпусов, а также обсуждаются перспективы их использования в различных областях современной науки о языке.

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

Финансирование. Благодарности.
Автор выражает свою признательность доц. Е.В. Зверевой за всестороннюю помощь в подготовке данной работы.

История статьи:
Дата поступления: 09.11.2020
Дата приема в печать: 08.01.2021

Для цитирования:
Чилингарян К.П. Корпусная лингвистика: теория vs методология // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2021. Т. 12. № 1. С. 196—218. doi: 10.22363/2313-2299-2021-12-1-196-218

UDK 81

Corpus Linguistics: Theory Vs Methodilogy

Kamo P. Chilingaryan

Peoples' Friendship University of Russia (RUDN University),
6, Miklukho-Maklaya str., Moscow, Russian Federation, 117198
Corresponding author: chilingaryan-kp@rudn.ru

Abstract. The article is devoted to a comprehensive study of the stages of formation and development of corpus linguistics. The purpose of the article is to analyze various scientific approaches to the scientific significance of this linguistic discipline and identify a set of concepts and criteria that form the foundation of this field. Corpus linguistics is one of the most promising and rapidly developing areas of language research. Linguistics of the XIX century set as its goal the study of language as such, and linguistics of the XXI century sees the relevance of the research not in identifying absolute linguistic categories and meanings but in the practical application of linguistic knowledge. The relevance of the article is determined by the fact that the linguistic corpus contains a vast potential, which the scientific community has not fully comprehended since the text as the main object of corpus linguistics in various forms of its implementation is one of the central components systems of language and speech-thinking activity of a modern native speaker of any language. The content and volume of linguistic corpora of various kinds allow obtaining reliable information about the modern and real use of a particular term: the corpus becomes a tool for analyzing the functioning of this term both in the linguistic field of morphology, syntax, and vocabulary and in the theory and practice of translation, identifying the register of its formal or informal usage. The fundamental novelty of this study's results allows us to speak about the legitimacy of the creation of corpus dictionaries and corpus grammars of a new generation, developed and verified concerning a specific fixed corpus. Simultaneously, the

author substantiates the proposition that the corpus nature of dictionaries and grammars increases their reliability and objectivity and avoids the subjectivity that is often characteristic of research-based solely on the intuition of a linguist. The corpus is a medium for obtaining new scientific data, the comprehension of which seems to be a priority for modern linguistic description and necessary in the scientific activity of a modern researcher. From our point of view, this article's relevance and novelty lie in the fact that the expediency of corpus research is an essential requirement of the time, associated with a new quality of linguistic reality and meeting the needs of modern society. The article examines the main stages of the formation of corpus linguistics as a scientific field, characterizes the scientific concepts and approaches inherent in each of these stages, provides an overview of the main conceptual provisions of corpus linguistics within the framework of domestic and foreign linguistics. The author analyzes in detail the polemics between representatives of various scientific directions and reveals the advantages of one or another approach, traces the similarities and differences between approaches to the study of corpora at various historical stages of their formation. The review's focus is the role and place of corpus studies of language in modern linguistics, comparison of the pro and contra arguments of the use of corpus technologies in linguistic description. Considerable attention is paid to the main criteria for the classification of corpora, a brief overview of the most famous corpora in history is offered, and the prospects for their use in various fields of modern language science are discussed.

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

Acknowledgement:
The author expresses his sincere gratitude to PhD Ekaterina V. Zvereva for her assistance and support in the article-writing process

Article history:
Received: 09.11.2020
Accepted: 08.01.2021

For citation:
Chilingaryan, K.P. (2021). Corpus Linguistics: Theory vs Methodology. *RUDN Journal of Language Studies, Semiotics and Semantics*, 12(1), 196—218. doi: 10.22363/2313-2299-2021-12-1-196-218

Корпусная лингвистика по-прежнему представляет только очень предварительные руководящие принципы теории, которая может установить связь отдельных текстов с текстовым корпусом, которая может использовать то, что является общим в корпусе, для идентификации типических характеристик языка, при этом лингвист может использовать выводы о часто повторяющихся паттернах для построения теории, которая объединяет рутинное (повседневное) использование языка и творческий подход лингвистического анализа».

M. Stubbs 1996. Text and corpus analysis.
Computer-assisted studies of language and culture.
Massachusetts: Blackwell.[1]

[1] Перевод К.П. Чилингарян.

Введение

История развития любого национального языка сопровождается меняющимися представлениями о его природе, сущности, строе, его общественной значимости и функционировании в глобальном масштабе. Такая ситуация актуальна и понятна, поскольку для эволюции научной лингвистической парадигмы в принципе свойственна смена точек зрения, подходов и аспектов изучения. Для лингвистики XIX века язык был интересен сам по себе, но для лингвистики XX и XXI веков стало актуально не столько теоретическое знание, сколько его прикладной аспект. В современной лингвистике текст, а затем и дискурс как объект исследования определил проблематику трудоемкости поиска научного материала, что потребовало оптимизации работы с языковым материалом. Как следствие, на пересечении собственно лингвистики и программирования возникло особое направление — корпусная лингвистика, приобретающая все большую популярность среди современных языковедов. В связи со стремительным развитием информационных технологий сбор и анализ практического материала и всего многообразия текстов на различных языках получает новое звучание и требует тщательно разработанных принципов и механизмов с учетом базовых положений корпусной лингвистики. Проф. Г.П. Мельников полагает, что любое исследование, осуществляемое лингвистом, должно быть ориентировано, по меньшей мере, на следующие этапы деятельности:

1) выбор принципов и оснований («эталонов») классификации изучаемых объектов;

2) процесс распределения объектов по классам в соответствии с этими основаниями («эталонами»);

3) осмысление, интерпретация, истолкование результатов распределения объектов по классам, объяснение причин такого распределения [1. С. 29].

Академик РАН В.А. Плунгян, заведующий кафедрой теоретической и прикладной лингвистики МГУ им. М.В. Ломоносова, характеризует корпусную лингвистику как «стремительное» и «суперсовременное» направление [2. С. 9]. A priori корпусная лингвистика обладает большим исследовательским потенциалом, однако, как показывают работы российских и зарубежных авторов, очевидны различия в подходах к созданию и использованию корпусов в России и за рубежом.

Автору статьи представляется, что на сегодняшний день корпусную лингвистику можно назвать одним из основных ресурсов изучения языка и языковой дескрипции. Если говорить о компьютерной лингвистике, то создание корпуса следует считать фундаментом данной лингвистической дисциплины, позволяющим создать автоматизированные приложения для обработки текстов и иных языковых и речевых проявлений [3. С. 147—227].

Обсуждение

Использование корпуса для исследования языковых явлений считается эмпирической методологией работы, основанной на использовании фактических

данных, образцов языкового и речевого узуса. Совокупность данных — это то, что обычно понимается как корпус в самом широком смысле этого слова. Например, Словарь Испанской Королевской Академии, основного «языковедческого» института Испании, определяет корпус следующим образом: «наиболее протяженный и упорядоченный набор данных или совокупность технических, литературных и прочих текстов, которые могут служить основой для исследования» [4]. «Толковый словарь русского языка» (под ред. Д.Н. Ушакова) определяет ***корпус*** как цельный свод каких-либо текстов [5].

Применение компьютерных технологий для сбора, упорядочивания и обработки данных явилось тем успешным фактором, который придал задаче создания корпусов современный облик, превратив отдельные опыты в стройную научную методологию и дисциплину, названную корпусной лингвистикой.

Наметим некоторые вехи, которые заложили фундамент корпусной лингвистики и способствовали развитию и консолидации ее как научного направления. Так, испанская исследовательница М. Вильяяндре Льямарес [6. С. 329—349] отмечает, что до XIX века корпус в лингвистике определялся как:

а) совокупность письменных текстов (данных);

б) объект, изучающийся в разрезе мертвых языков (латынь, санскрит);

с) объект, к которому лингвисты могли приблизиться только с помощью метода корпуса, поскольку не располагали возможностью сбора языковых данных посредством живых носителей.

В XIX веке и до середины XX века эта методология исследования продолжала применяться и основывалась на сборе большого количества текстов для:

1) объяснения процесса овладения языком детьми (транскрипция интеракции ребенка с родителями);

2) установления орфографических норм;

3) составления лексических списков для преподавания вторых языков;

4) осуществления компаративных исследований языков;

5) разработки дескриптивных грамматик.

Данные направления исследований отмечают такие авторы, как McEnery [7. С. 448—463]; McEnery, Wilson [8. С. 13—176]; McEnery, Wilson [9. С. 103—106]; McEnery, Xiao, Tono [10].

В первой половине XX века американская структурная лингвистика заложила основы корпусной лингвистики как эмпирической методологии, основанной на наблюдении за языковыми данными, хотя сам термин «корпусная лингвистика» появился значительно позже, в начале 80-х годов XX века. Исследователи того периода полагали, что корпус является единственным инструментом, подходящим для изучения языков, утверждая, что сам корпус может предоставить необходимые данные для исчерпывающей дескрипции того или иного языка. Эта новая концепция корпуса, т.н. «структурный корпус», характеризовалась следующими признаками:

1) совокупностью устных образцов или письменных транскрипций (данные);

2) определением цели — изучение живых языков, но тех, которые ранее не были письменно задокументированы (языки американских индейцев);

3) необходимостью, так как сбор устных речевых образцов был единственной возможностью «доступа» к этим языкам;

4) фокусированием работы на фонетических и (морфо)фонологических аспектах как тех уровнях, на которых возможно провести «инвентаризацию» всех элементов, принимая во внимание их законченную природу;

5) игнорированием фактора репрезентативности результатов: поскольку анализ данных осуществлялся визуально и вручную, то отсутствовала возможность оперировать большим количеством данных (именно поэтому данная методология подвергалась критике и признавалась необъективной) [6. С. 329—349].

Многие лингвисты, в частности Н. Хомский [11. С. 553—58; 12. С. 41; 153; 171], исповедовавший рационализм в языковых исследованиях, способствовали тому, что эмпирическая методология корпуса американских структуралистов в 1960—70-е годы уступила место иному подходу: так называемому рационализму, по Н. Хомскому, «интуиции лингвиста» [11. С. 558]. Н. Хомский критиковал корпусную лингвистику с теоретической точки зрения. Он считал, что в языковых исследованиях единственным ресурсом лингвиста является его интуиция, которая представляет собой единственный значимый критерий; синтаксис должен являться основным объектом исследования [Там же].

Различие концептов начального периода корпусной лингвистики и теорий Н. Хомского можно представить в следующей последовательности:

Корпусная лингвистика:	Н. Хомский:
1) фокусируется на фонетике и фонологии	1) фокусируется на синтаксисе
2) язык воспринимается как законченное явление	2) язык — это объект, не имеющий границ
3) корпус способен объяснить все содержащиеся в самом себе явления	3) интуиция лингвиста как единственный способ дескрипции
4) корпус является полным и совершенным	4) корпус — незавершенная совокупность

Наряду с критическим анализом теоретических идей Н. Хомского, ряд исследователей также отмечали практические проблемы в первых опытах корпусной лингвистики. Обработка данных была крайне медленной, дорогой и нередко ошибочной. Так, Д. Аберкромби называл корпусные исследования «псевдотехниками», способствующими недостоверности аналитики [13. С. 22].

Именно появление компьютерных технологий дало новый импульс этому научному направлению. Некоторые исследователи называют этот период корпусной лингвистикой нового поколения [6. С. 340]. Основными характеристиками данного периода исследования корпусов (60—70-е годы) были:

а) наличие компьютеров: только в этот период компьютеры стали обладать достаточной мощностью для анализа данных (хотя уже в конце 40-х годов Р. Бусса проводит первые эксперименты по компьютерной обработке данных корпусов, ссылаясь на Т. Макэнери) [7. С. 451];

б) репрезентативный характер данных: большинство проектов были направлены на сбор письменных текстов, анализ которых позволил бы характеризовать состояние языка в данный период.

В 50-е годы А. Жуланд, основываясь на работу Т. Макэненри [7. С. 459], установил рамочные критерии для языковых образцов:

а) репрезентативность и сбалансированность;

б) тенденция к неиспользованию образцов устной речи, объясняющаяся трудностями технического плана и сложностями транскрибирования, таким образом доминируют корпусы письменных текстов;

в) размер: миллион слов [14. С. 320—339].

Обосновывая необходимость репрезентативности корпуса, Д. Байбер отмечает, что если понятие «общий язык» представляет собой абстрактную категорию, а язык функционирует как система различных жанров и/или стилей, то референтный корпус должен включать все стили и жанры речи, а также территориальные диалекты. Говоря о социальной представленности языка, Д. Байбер утверждает, что в корпусах необходимо фиксировать диалекты, социолекты и профессиональные языки, или языки для специальных целей. Д. Байбер уточняет [15. С. 209—213], что язык должен быть представлен в историческом ракурсе, то есть включать тексты всех известных исторических эпох. Таким образом, Д. Байбер считает, что репрезентативность корпуса связана со сбалансированностью, пропорциональной представленностью жанров и стилей языка всех слоев общества, которая соответствует существующей в реальности [16. С. 243—257]. В целом репрезентативность рассматривается Д. Байбером как представленность в корпусе текстов широкого спектра функциональных стилей и жанров. В свою очередь, П. Бейкер пишет, что понятие репрезентативности тесно связано с понятием валидности или соответствием полученных данных реальному состоянию языка в данной сфере употребления [17. С 30—35; 160—169], при этом, как полагают исследователи, полная репрезентативность в корпусах недостижима и невозможна. Как отмечает Е.А. Красина, словесный образ, сложные образные структуры управляются глубинными ассоциативными связями, которые обнаруживаются на различных уровнях динамической структуры художественного текста [18. С. 858].

Напомним, что первые значимые корпуса появились на базе английского языка в 1960-е годы. В 1959 г. в Великобритании Р. Кирк (University College, London, GB) заложил теоретическую основу для создания ‘Survey of English Usage Corpus’ (SEU), первого европейского проекта корпуса, созданного для дескрипции и анализа языка. Корпус состоял из 200 текстов по 5000 слов, а сбор материала начался в 1961 г. и представлял собой попытку систематизированного описания британского варианта английского языка 1955—1985 гг. на основе транскрипций записей устных высказываний и письменных

текстов. Данный проект наметил основные нормы и процедуры будущей корпусной лингвистики.

В 1963 г. в США У.Н. Френсис и Г. Кучера создали первый корпус текстов на электронном носителе ‘Brown University Corpus of American English’ (‘Brown Corpus’). В него вошли 500 текстов (по 2 000 слов в каждом) пяти самых популярных жанров англоязычной прозы США. Тексты были отобраны из различных публикаций 1961 г. К корпусу прилагались указатель частотности и алфавитно-частотный указатель, а также некоторые статистические распределения. Целью корпуса, помимо описания американского варианта английского языка на примере письменных текстов, было определение стандарта для развития компаративных исследований, основывающихся не только на подобных «корпусах» текстов английского языка. Компьютерный анализ данного корпуса выявил такие известные сегодня закономерности, как, например: наиболее частотными словами являются те, которые не несут смысловой нагрузки, такие как ‘the’, ‘of’; или большинство слов, имеющих определенное лексическое значение, появляются в корпусе лишь единожды и др. Помимо «лексико-статистического» вклада данный проект послужил базой для создания первого словаря, основанного на корпусе: American Heritage Dictionary (1969). Также упомянем ‘Lancaster-Oslo/Bergen Corpus’ (LOB), корпус, созданный при участии Дж. Лича (университет Ланкастера, Великобритания), С. Йохансона (Университет Осло, Норвегия) и ‘Norwegian Computing Centre for the Humanities’(Берген). Корпус состоял из миллиона слов, представляющих образцы британской английской письменной речи 1961 г. Корпус был составлен согласно критериям Брауновского корпуса и представлял собой его британский эквивалент.

В 1975 г. Й. Свартвик [19. С. 7] (Университет Линда, Швеция) начал работу по созданию информационной базы для классификации устных образцов, не транскрибированных в ‘Survey of English Usage Corpus’ (SEU), проект получил название ‘Survey of Spoken English’ (SSE), и в результате был создан London-Lund Corpus of Spoken English (LLC), состоящий из 500 тыс. слов британского варианта английского языка в разговорном регистре, записанных в период 1953—1987 гг. Записи транскрибируются орфографически и дополняются информацией о просодии и паралингвистическими характеристиками. Подобный проект до сих пор не был повторён. Все корпусы этого периода имели структурированный и репрезентативный формат, которого нередко не хватает современным корпусам.

Корпусная лингвистика как самостоятельная дисциплина окончательно оформилась в 90-х гг. XX века. Именно в этот период электронные корпусы превращаются в обязательный ресурс исследования языка, создания лингвистических гипотез и построения систем обработки данных естественного языка. На «возрождение» корпусной лингвистики (КЛ), по нашему мнению, оказали большое влияние ряд ученых, среди которых хотелось бы отметить Дж. Лича [20. С. 105—122], который начал полемику с критикой теоретического и практического толка идей Н. Хомского и Д. Аберкромби (см. выше).

Если в 1960-е годы XX столетия данная критика была отчасти объективной, то сейчас, по словам Дж. Лича, благодаря эволюции компьютерных технологий, в защиту создания корпусов приводятся следующие основные аргументы:

1) корпус является научной методологией и поэтому имеет безусловное преимущество по сравнению с интуицией, так как может быть контролируемым и не учитывать образцов, изобретенных лингвистами, заинтересованными в прогнозируемом результате. К тому же в области количественных данных, таких, как, например, частотность, интуиция является неприемлемым инструментом — наша перцепция частотности абсолютна субъективна;

2) грамматический характер текстов корпуса, в связи с чем корпус отражает языковую компетенцию (при этом Н. Хомский заявлял, что поскольку корпусы являются образцами речевых проявлений, образцами использования языка, то они не отражают языковую компетенцию. Однако работа В. Лабова [21. С. 1—44] доказали высокий процент грамматических последовательностей в корпусе);

3) значимость количественных данных: корпусы представляют собой ни с чем не сравнимый источник получения этого рода данных;

4) если структура корпуса является научно выверенной, то данные, относящиеся к частотности узуса, будут репрезентативными для всего языка в его целостности;

5) использование компьютера опровергает утверждение об использовании «псевдонаучных» техник;

6) компьютерная обработка значительных массивов информации при низкой стоимости, высокой скорости позволяет избежать субъективных человеческих ошибок [21. С. 105—122].

Российские ученые, например, один из руководителей проекта создания «Национального корпуса русского языка» акад. В.Н. Плунгян, поддерживают концепцию лингвистической значимости корпуса: «корпус необходим исследователям, занимающимся систематизацией фактов об анализируемом языке, а также в академических целях, поскольку таким образом процесс освоения языковых компетенций происходит быстрее» [2. С. 11]. Над национальным корпусом русского языка работали и продолжают работать такие ученые, как Г.И. Кустовая, А.Е. Поляков и Д.В. Сичинава и мн. другие. Из советских, ныне российских ученых, работающих в области корпусной лингвистики, можно также отметить труды Е.А. Красиной и М.Л. Новиковой [18, 22], В.Н. Денисенко и Н.В. Перфильевой [23].

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, корпусы 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 (тем) (например, медицинские или юридические тексты) [25; 26. С. 71—79].

К этому периоду 1990-х гг. относится и создание научного понятийного аппарата. Дж. Синклер определяет корпус как набор текстов на естественном языке, выбранных для характеристики разнообразия языков [27. С. 171]. Приведенная дефиниция подчеркивает основной критерий для создания корпуса: естественный, то есть первозданный необработанный текст в устной или письменной форме, естественное речевое проявление языковой формы. Позже понятие «корпус» получает расширение. М. Стаббс считает, что корпус — это собрание текстов, предназначенное для какой-то цели, обычно для исследований или обучения. Корпус — это не то, что делает или знает говорящий, но что-либо, выстроенное исследователем. Это регистрация совокупной деятельности значительного количества пользователей языка, структурированная под цели изучения и создаваемая для выявления характеристик наиболее типичного использования языка [28. С. 239—240]. М. Стаббс полагал, что компьютерное исследование обширных корпусов может показать путь выхода из парадоксов дуализма языка. Ряд исследователей вели дискуссии о том, следует ли считать корпусную лингвистику теорией или исключительно методологией [8. С. 25; 26]. Так, Р. Симпсон и Дж. Свалес называют корпусную лингвистику техникой или технологией создания и анализа

корпуса [30. С. 1—14]. Большинство исследователей приходят к выводу, что рассматриваемая нами дисциплина языковедения представляет собой реализацию эмпирического подхода к наблюдаемым данным, сохраняемым в виде электронных корпусов, своего рода методологический инструмент для изучения языков, предоставляющий инновационные возможности для дескрипции, анализа и преподавания языков. Также корпусная лингвистика является эмпирическим фундаментом для создания учебных и методических материалов, таких как грамматики, словари и т. п. как на основании дискурсов общего плана, так и специализированных, имеющих устную или письменную фиксацию. Так, чилийский ученый Г. Пароди [31. С. 95] полагает, что корпусная лингвистика — это совокупность методологических принципов для изучения любой лингвистической области, обслуживающая цели исследования языка в его узусе исходя из материала лингвистических корпусов и опираясь на компьютерные технологии и программы ad hoc. Поэтому нельзя трактовать корпусную лингвистику как сферу лингвистики, подобную фонологии, семантике, синтаксису, но как метод исследования, применимый во всех дисциплинах лингвистики, на всех уровнях языка и с точки зрения различных теоретических подходов. Приведем еще ряд определений корпусной лингвистики как «области лингвистики, специализирующейся на получении результатов от изучения корпусов» [32. С. 61—90]; «изучения языка на базисе текстового корпуса» [33. С. 1]; «использования обширной коллекции доступных текстов в обработанном компьютере виде» [19. С. 7]; «набора текстов, которые, как предполагается, являются репрезентативными для данного языка, диалекта или другого говора языка, который будет использоваться для лингвистического анализа» [34. С. 17]; «набора выбранных и упорядоченных фрагментов языка в соответствии с явными лингвистическими критериями для использования в качестве образца языка» [35. С. 14]; «набора машиночитаемых текстов конечного размера, отобранный для максимальной репрезентативности рассматриваемого языкового разнообразия» [35 С. 14; 8. С. 24]; как то: «корпус — это образец языка, который выстраивается на основании селекции текстов, сделанной согласно детерминированным критериям и цели исследования» [36. С. 151]. При этом «термин „корпус“» должен правильно применяться только к хорошо организованному сбору данных, собранных в рамках структуры выборки, предназначенной для изучения определенной лингвистической характеристики (или набора характеристик) через собранные данные» [7. С. 449]; «корпус — это совокупность текстов естественного языка, собранных в однородном электронном формате, отобранных и упорядоченных в соответствии с эксплицитными критериями, служащая в качестве модели современного или диахронного состояния или уровня определенного языка для научного исследования или прикладной деятельности, связанной с лингвистическим анализом» [37. С. 45—46]; «термин „корпус“», используемый в современной лингвистике, лучше всего можно определить как набор отобранных текстов, письменных или устных, в машиночитаемой форме, которые

могут быть аннотированы различными формами лингвистической информации» [10. С. 4].

Что касается узуальности языка, то, по мнению Е.А. Красиной [22], при большей развернутости латинизма (высказывание-цитата), он обрастает узуальными смыслами. И наоборот, при десемантизированном содержании латинизма он тяготеет к знакам-индексам и выполняет функции текстовых скреп и дискурсивных маркеров.

Если говорить об ученых, представляющих отечественное языкознание, то здесь мы тоже видим ряд непротиворечащих друг другу определений понятия «корпус». Так, В.П. Захаров считает, что корпус — это большой, представленный в электронном виде, структурированный и размеченный, филологически представительный массив языковых данных, предназначенных для решения определенных лингвистических задач [38. С. 3]. Н.В. Козлова определяет корпус как собрание текстов одного или нескольких языков, связанных между собой определенными параметрами, как собрание письменных и устных высказываний. При этом составные части корпуса, тексты, состоят из данных, а также, возможно, из метаданных, описывающих эти данные, и из лингвистических аннотаций, которые эти данные упорядочивают [39. С. 80].

Итак, обобщив различные определения, можно выделить некоторые общие критерии:

а) тексты должны содержаться на электронном носителе или в сети Интернет;

б) размер корпусов должен прогрессивно наращиваться, достигая 100 миллионов слов, хотя для специальных целей могут создаваться корпусы и меньших размеров (отметим, что ранее существовала точка зрения, что чем больше корпус, тем больше возможностей у исследователя отразить реальное функционирование языка во всей его вариативности, в настоящее время в приоритете исследователей конкретная направленность корпуса (корпус языка Сервантеса и т.д.);

в) открытый характер: корпус постоянно актуализируется, т.н. корпус монитор;

г) аутентичность данных: тексты должны быть реальными образцами использования языка, являющегося объектом исследования;

д) критерии отбора: тексты должны выбираться не произвольно, а в соответствии с лингвистическими и/или экстралингвистическими задачами, которые корпус преследует. Именно это является основным критерием, отличающим корпус от прочих собраний текстов, таких как архивы или электронные библиотеки;

е) репрезентативность: селекция текстов должна отвечать статистическим параметрам, которые будут гарантировать, что текст представляет ту разновидность языка, которая является объектом исследования (репрезентативный образец). Эта разновидность языка может относиться к произведению определенного автора, к определенному историческому периоду, жанру и т. д.;

ж) коммерческая направленность: корпусы не являются достоянием и прерогативой исключительно исследовательских центров, многие проекты могут реализовываться в рамках деятельности коммерческих структур, таких как издательские концерны;

з) расширение репертуара языков, для которых создаются корпусы, а также создание мультилингвальных корпусов;

и) расширение ареала исследования различных языковых аспектов — от грамматических до дискурсивных, при этом широко должны учитываться исторические, психолингвистические и культурологические факторы;

к) представленные языковые данные должны иметь определенную разметку для лингвистического анализа (под разметкой В.П. Захаров понимает приписывание текстам и их компонентам специальных меток: внешних, экстралингвистических, структурных и собственно лингвистических, описывающих лексические, грамматические и прочие характеристики элементов текста [38. С. 6];

л) проведенный анализ предполагает возможность классификации полученного материала с учетом тематики текста, степени специализированности, жанровой характеристики и др.

Очевидно, что одним из важнейших критериев является доступность корпуса в электронном виде. Вторым важным моментом выступает, безусловно, его репрезентативность. Как считает А.Е. Кибрик, репрезентативность можно оценить «по изменению относительной частоты рассматриваемого явления при увеличении выборки. Если относительная частота явления от прибавления каждого последующего фрагмента текста будет изменяться все меньше и меньше, то это означает, что корпус в целом репрезентативен» [40. С. 21]. Мы полагаем, что репрезентативность корпуса является тем обязательным условием, которое позволяет придать набору различных текстов статус корпуса текстов, позволяющего осуществить лингвистический анализ. Вместе с тем вопрос о безусловной репрезентативности того или иного корпуса следует признать до сих пор открытым, поскольку языковая деятельность человеческого общества характеризуется чрезвычайным разнообразием, что затрудняет возможность объективного отражения в корпусе всех языковых вариантов, национально-культурных вариантов языка. Это, на наш взгляд, объясняет некоторую субъективность результатов анализа корпусов текстов.

Обратимся к примерам наиболее известных и значимых корпусов на сегодняшний день:

1. ‘The Bank of English’ — более 524 миллионов слов, образцы речевых употреблений в устной и письменной форме из различных национальных вариантов английского языка: британский, американский, канадский и австралийский). Тексты размечены согласно грамматическим категориям, более 200 миллионов слов были проанализированы с точки зрения синтаксиса. Важной характеристикой является непрерывная актуализация корпуса. В настоящее время проект носит название Project COBUILD и реализуется на базе Университета Бирмингема под руководством уже упоминаемого нами Дж. Синклера

в сотрудничестве с издательством Collins COBUILD. Корпус был инициирован в 1991 г., но с 1980 г. COBUILD уже собирал электронные тексты для своих словарей. Все данные находятся в открытом доступе на странице 'Collins Word Web', база данных составляет более двух с половиной миллиардов слов, к которым каждый месяц добавляется 35 миллионов. Это самый обширный ресурс подобного типа в мире.

Еще один важный корпусный ресурс для английского языка — это «British National Corpus» (BNC), насчитывающий 300 миллионов слов современного британского английского языка в письменной и устной форме. Проект реализуется под эгидой научно-промышленного консорциума во главе с Оксфордским издательством Oxford University Press вместе с другими издательствами, специализирующимися на словарях, Университетом Ланкастера, Оксфордским университетом и Британской библиотекой. Создание корпуса велось в 1991—1994 гг., это пример корпуса закрытого характера, целью которого являются исследования в области британского варианта английского языка конца 20 века для разработки справочных материалов (состоит из 90% письменных текстов и 10% устных текстов).

2. 'Corpus de Referencia del Español Contemporáneo' (CREA), банк данных современного испанского языка (с 1975 г. по настоящее время), разработанный Испанской Королевской Академией (la Real Academia Española). Насчитывает более 200 миллионов слов. 90% образцов составляют письменные тексты, остальное — устные. При отборе образцов корпус учитывает географические, тематические, хронологические критерии, а также источник, откуда текст был получен. Банк данных считается корпусом-монитором — периодически добавляются новые тексты с целью повышения репрезентативнсти корпуса. Это самый значимый корпус испанского языка, служащий как для академических исследований, так и для создания коммерческих продуктов.

3. 'Corpus Diacrónico del Español' (CORDE), банк данных испанского языка в диахроническом аспекте, собрание текстов начиная с периода образования испанского языка до 1975 г. Этот корпус является историческим дополнением CREA, насчитывая в совокупности более полумиллиона слов.

Важным аспектом в теории изучения лингвистических корпусов можно считать также их различные типы. Н.В. Козлова считает, что все существующее множество корпусов текстов можно разделить на три категории: 1) находящиеся в свободном доступе; 2) находящиеся в частичном доступе и 3) коммерческие [39. С. 76—89]. Дж. Синклер [35] и Дж. Торруэлла и Дж. Листери [41. С. 45—77] разработали параметры классификации текстов (отметим, что на практике эта типология не всегда оказывается эксплицитной):

1) разновидность языка;

2) количество языков, к которым относятся тексты;

3) границы корпуса (корпус, чья цель — описание подъязыка (юридический, язык информатики и т. д.) может иметь ограниченный размер);

4) общий или специализированный характер текстов;

5) временной срез, который тексты охватывают;

6) способы анализа и обработки данных, применяемые к корпусу.

В большинстве случаев эти критерии определяются целью, для которой предназначен корпус: исследование авторских произведений (например, поэмы Г. Додохяна) или литературные произведения определенного исторического периода, описание национального языка в целом (современный армянский язык) или определенная разновидность языка, территориальный диалект, язык для специальных целей, или конкретный лингвистический аспект (напр., культурная норма Еревана, медицинские тексты, тексты рекламы и т.д.), получение определенного коммерческого продукта (например, туристический разговорник).

Поскольку разнообразие существующих корпусов «определяется многообразием исследовательских и прикладных задач, для решения которых они создаются» [38. С. 12], важным аспектом понятийного аппарата корпусной лингвистики является вопрос о типологии корпусов.

По критерию формы языка можно выделить следующие типы корпуса — письменные, устные и смешанные. Большинство из существующих корпусов относятся к письменным, например, Корпус испанского языка (исторический) [URL: https://www.corpusdelespanol.org; дата обращения: 02.11.2020]; создан в Иллинойском университете (США) и содержит 100 миллионов слов, включая письменные тексты XIII—XX вв.), либо смешанным, при этом предпочтении отдается письменным образцам как в уже упомянутом нами корпусе CREA.

Под устным корпусом подразумевается структурированная совокупность речевых фрагментов, которая обеспечена программными средствами доступа к ним [42. С. 81—84]. В устных корпусах образцы речи представляют собой либо орфографические транскрипции аудиозаписей, либо непосредственно являются аудиозаписями, сопровождающимися орфографическими и/или фонетическими транскрипциями. В качестве примера приведем Корпус армянского языка, который состоит из 20 рассказов и 20 пересказов десяти носителей армянского языка. Записи производились в Ереване в 2004—2005 гг. Все информанты являются ереванцами, возраст информантов на момент записи — от 17 до 25 лет. Общая длительность звучания — около 42 минут; объем корпуса — около 4,5 тысяч словоупотреблений. Для корпуса выполнена модифицированная минимальная транскрипция, включающая в себя армянскую графику, латинизированную запись, поморфемную нотацию, глоссирование [URL: https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf; дата обращения: 01.11.2020] и русский перевод. Подобные корпусы записываются при соблюдении строгих условий и обычно состоят из отдельных фрагментов, имея ограниченный объем, позволяющий не использовать большое количество информантов. Целью создания таких корпусов в большинстве случаев является не столько анализ фонетических и просодических характеристик, сколько социолингвистический анализ или дискурсивный анализ (например, проект PRESEA для испанского языка Испании и Латинской Америки). PRESEEA — это реализуемый с начала нового тысячелетия проект по созданию корпуса разговорного испанского языка, представляющего

испаноязычный мир в его географическом и социальном разнообразии. Эти материалы собираются с учетом социолингвистического разнообразия испаноязычных сообществ. PRESEEA объединяет около 40 групп социолингвистических исследований и проводится под эгидой Испанской Королевской Академии [43].

Следующим критерием для классификации корпусов является количество используемых в них языков: одноязычные, двуязычные и многоязычные. В зависимости от специфики текстов среди одноязычных корпусов можно выделить две разновидности:

1) корпусы, охватывающие весь язык, т.н. общий корпус, предназначенный для отражения языкового или лингвистического разнообразия максимально сбалансированным образом: чем больше типов текстов, их разновидностей (устные и письменные), жанров и тематики, тем более качественным признается корпус;

2) корпусы, охватывающие только язык для специальных целей. Корпусы для специальных целей чаще всего соотносятся с определенным типом дискурса и способны учитывать различные лингвокультурологические феномены в процессе коммуникации. Критерием репрезентативности в данном случае является максимально возможное объективное представление какого-либо явления, определенное создателями данного корпуса. Так, например, корпус англоязычных пословиц, отражающий использование в речи носителей языка определенного периода времени и географического региона, не будет релевантным при изучении английской политической метафоры [44. С. 126].

Специализированные корпусы собирают тексты, которые могут предоставить данные для описания определенного типа языка («подъязыка»). Например, корпус, который «собирает» только поэтические или юридические тексты. В качестве примера приведем ‘Corpus textual especializado plurilingüe’ (Многоязычный специализированный текстовый корпус), разработанный испанским Институтом прикладной лингвистики Университета Помпеу Фабра в 1993 г. Он состоит из текстов на каталонском, испанском, английском, французском и немецком языках и охватывает такую тематику, как экономика, право, окружающая среда, медицина и информатика. Корпус создан с целью изучения особенностей функционирования языка в каждой из конкретных областей и выборки материала для последующего создания специализированных словарей, тезаурусов и т.д.

Двуязычные и многоязычные корпусы могут состоять из текстов, которые не обязательно являются переводными вариантами друг друга и не были выбраны для корпуса согласно унифицированным критериям, то есть тексты могут быть представлены соизмеримо или параллельно. Когда корпус состоит из выборки текстов на более чем одном языке, схожих по своим характеристикам и общим принципам отбора, можно говорить о «сопоставимом корпусе» или «парных текстах». Они используются в основном для сравнения языковых разновидностей в сравнительных исследованиях. Самым известным из них является ‘The International Corpus of English’ (ICE), с 1990 года

фиксирующий письменные и устные материалы из различных разновидностей английского языка в период после 1989 года. В настоящее время к проекту подключены 19 стран, в которых английский язык имеет статус официального. Каждый корпус состоит из миллиона слов, и все они имеют одинаковую структуру и перечень аннотаций. В 1992 г. была создана Европейская корпусная инициатива European Corpus Initiative (ECI) — международная организация, занимающаяся созданием обширного многоязычного корпуса для научных целей. В данном соизмеримом корпусе содержатся в основном тексты на европейских языках, а также тексты на таких языках, как турецкий, китайский, японский и др., общим объемом более 98 млн слов.

Когда корпус содержит тексты на нескольких языках, но, в отличие от сопоставимого корпуса, это одни и те же тексты и их переводы или эквиваленты на одном или нескольких языках, используется термин *параллельный корпус* или *bitexts*. Подобные корпусы особенно полезны при изучении переводческих практик и технологий, для сопоставительного анализа текстов «оригинал — перевод» в целях обучения методам и приемам перевода, а также в двуязычной или многоязычной среде таких международных организаций, как ООН, НАТО, или в таких практически билингвальных государствах, как США или Канада. В качестве примера можно обратиться к «Параллельному корпусу CLUVI» Университета Виго или European Parliament Proceedings Parallel Corpus 1996—2011, где представлены параллельные тексты заседания Европейского парламента на всех европейских языках с переводом на английский.

В зависимости от периода времени, охватываемого текстами, корпусы подразделяются на:

А) диахронический, или исторический корпус: включает тексты разных временных срезов, позволяющие анализировать эволюцию языка в течение длительного периода и исследовать историческое развитие какого-либо языкового явления, либо всей языковой системы в целом, что отличает их от корпуса — монитора, который не охватывает такие длительные периоды времени (упоминаемый нами корпус CORDE);

Б) синхронный корпус: представление текстового материала для рассмотрения состояния языка как системы в определенный момент времени (Британский национальный корпус).

Также существует возможность классифицировать корпус согласно существующей в них разметке, а именно как неразмеченные и размеченные. Неразмеченный корпус — это массив текстов, которые содержат определенное количество упоминаний искомого элемента. При этом результаты поиска, предоставляемые в неразмеченных корпусах, могут быть использованы в лингвистических исследованиях, но только с чисто статистической точки зрения. Размеченные с точки зрения морфологических, синтаксических, просодических и иных характеристик корпусы предоставляют намного больше возможностей для проведения лингвистического анализа.

Безусловно, данный список перечисленных возможных типологических критериев не является ни закрытым, ни претендует на строгие границы разграничения типов корпуса.

Заключение

Таким образом, корпус, являющийся отражением языка, представляет собой, равно как и сам язык, динамически развивающаяся система, предполагает все новые критерии и подходы к описанию и анализу языкового материала и разработку новых методологических процедур. Корпус может предоставить подробную информацию о конкретном языке, но невозможно собрать корпус, охватывающий весь язык, т. к. невозможно собрать все образцы использования этого языка, поэтому всегда следует исходить из того, что корпус — это всего лишь некоторое конечное собрание образцов бесконечной вселенной языка.

Можно сделать вывод о том, что корпус — это представленный в электронном виде, как правило, размеченный для анализа в лингвистических целях, обеспеченный сравнительно простой в использовании поисковой системой репрезентативный массив неотредактированных текстов, репрезентирующих максимальное множество вариантов языка. Если в период зарождения корпусной лингвистики исследователи указывали, что можно пренебречь языковой вариативностью, то с появлением электронных корпусов многообразие форм существования языка стало более наглядным и возможности исследования языковых данных расширились. Обширная типология корпусов, созданных с учетом различных критериев, и их многообразие позволяет и лингвисту, и непрофессиональному пользователю выбрать тот, который соответствует целям и задачам определенного и конкретного научного исследования.

В настоящее время корпус является уникальным ресурсом для любого лингвистического исследования в целом и компьютерной лингвистики в частности. Его основные преимущества заключаются в том, что он состоит из реальных образцов языка, обеспечивает объективность полученных результатов и выводов, а также дает возможность достаточно просто верифицировать достоверность той или иной теории, ее достоинства и недостатки. Благодаря внедрению компьютеров с постоянно увеличивающейся емкостью хранения и скоростью обработки данных доступ к языковым и речевым образцам стал быстрым и надежным, как и извлечение данных, их обработка и анализ. С другой стороны, корпус позволяет получать статистические и количественные данные, которые, в ином случае, были бы недоступны ввиду высокой стоимости или невозможны ввиду недостоверности результатов, полученных при обработке информации «ручным способом» с учетом значительных размеров отдельных корпусов. Благодаря разработке методологии и методов корпусной лингвистики исследователям и любым другим пользователям доступны дескриптивные исследования языков, поддерживаемые корпусом на любом из лингвистических уровней: это фонетико-фонологический, грам-

матический, семантический, прагматический и др. Исключительная значимость применения корпусов в качестве источника данных обнаруживается в процессе обучения родному и иностранным языкам или разработки дидактических материалов, словарей, грамматик, прочих продуктов, связанных с машинным переводом или речевыми технологиями и проч.

Изучив некоторые условия «бэкграунда» корпусной лингвистики, а также современное состояние этой научной дисциплины, проанализировав имеющиеся на сегодняшний день данные, как и продолжающие создаваться и пополняться корпусы, можно сделать вывод о том, что эволюция, динамика этого перспективного научного направления актуальна для лингвистической теории и практики, и, по словам Г. Пароди, достигает точки кипения, однако механизмы компьютерной лингвистики переживают процесс постоянных изменений и корректировок, позволяя существенно обогатить и сделать более комплексными и достоверными результаты лингвистических исследований.

Библиографический список

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/ushakov/корпус (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. *McEnery T., Wilson, A.* Corpus Linguistics. Edinburg: Edinburgh University Press, 1996.
9. *McEnery T., Wilson A.* Corpus Linguistics. Edinburgh: Edinburgh University Press, 2001.
10. *McEnery 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.). Dordrecth: D. Reidel, 1969. P. 53—68.
12. *Chomsky N.* Language and mind. Cambridge, 2006.
13. *Abercrombie D.* Studies in Phonetics and Linguistics, London: Oxford University Press, 1965. Режим доступа: http://www.davidcrystal.com/Files/BooksAndArticles/-4896.pdf (дата обращения: 06.11.2020).
14. *Juilland A.G., Brodin D.R., Davidovitch C.* Frequency dictionary of French words. Hague—Paris: Mouton, 1970.
15. *Biber D., Conrad S., Reppen R.* Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press, 1993.
16. *Biber D.* Representativeness in corpus design // Literary and Linguistic computing. 1993. Vol. 8(4). P. 243—257.

17. *Baker P., Hardie A., McEnery T.* Glossary of Corpus Linguistics. Edinburgh: Edinburgh University Press, 2006.
18. *Красина Е.А., Новикова М.Л.* Феномен языка в парадигмах функциональной семантики и лингвосемиотики (V Новиковские чтения. Москва, 18—19 апреля 2019 г.) // Russian Journal of Linguistics. 2019. Т. 23. № 3. С. 856—864. DOI: 10.22363/2312-9182-2019-23-3-856-864.
19. *Svartvik J.* Corpus linguistics comes of age // Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82 (Stockholm, 4—8 August, 1991) / J. Svartvik (ed.). Berlin—New York: Mouton de Gruyter, 1992. P. 7—13.
20. *Leech G.* Corpora and theories of linguistic performance // Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82 (Stockholm, 4—8 August, 1991) / J. Svartvik (ed.). Berlin—New York: Mouton de Gruyter, 1992. P. 105—122.
21. *Labov W.* The logic of non-standard English // Georgetown Monographs on Language and Linguistics. 1969. № 22.
22. *Красина Е.А., Перфильева Н.В.* Семантические параметры квантитативных единиц в разноструктурных языках // Вопросы когнитивной лингвистики. 2018. № 1(54). С. 126—136. DOI: 10.20916/1812-3228-2018-1-126-136.
23. *Денисенко В.Н., Красина Е.А*. *Перфильева Н.В.* Принцип двойного означивания в языке и слове // Вопросы когнитивной лингвистики. 2016. № 3(48). С. 103—108.
24. *Aarts J., Meij, W.* (eds.) Corpus Linguistics. Amsterdam: Rodopi, 1984.
25. Manual for the Corpus of Early English Correspondence Sampler CEECS / Nurmi A. (ed.). Helsinki, 1998.
26. *Taavitsainen I., Pahta P.* Corpus of Early English Medical Writing // Computers in English Linguistics. 1997. № 21. P. 71—79.
27. *Sinclair J.* Corpus, Concordance, Collocation. Oxford, 1991.
28. *Stubbs M.* Words and phrases: corpus studies of lexical semantics. Oxford: Blackwell, 2001.
29. *Stubbs M.* Corpus analysis: the state of the art and three types of unanswered question // System and corpus: Exploring connections / Hunston S., Thompson G. (eds.). Londres: Equinox, 2006. P. 15—36.
30. *Simpson R., Swales J.* Introduction to North American perspective on corpus linguistics at the millennium // Corpus linguistics in North America. Selections from the 1999 Symposium. Ann Arbor: The University of Michigan Press, 2001. P. 1—14.
31. *Parodi G.* Lingüística de Corpus: Una introducción al ámbito // Revista de Lingüística Teórica y Aplicada. 2008. № 46(1). P. 93—119.
32. *Abaitua J.* Tratamiento de corpora bilingües // Tratamiento del lenguaje natural: tecnología de la lengua oral y escrita / M.A. Martí, J. Llisterri (eds.). Soria—Barcelona: Fundación Duques de Soria/Edicions de la Universitat de Barcelona, 2002. P. 61—90.
33. *Aijmer K., Altenberg B.* (eds.) English Corpus Linguistics. Studies in Honour of Jan Svartvik. London: Longman, 1991.
34. *Francis W.N.* Language Corpora B.C. // Directions in Linguistics: Proceedings of Nobel Symposium 82 (Stockholm, 4—8 August, 1991) / J. Svartvik (ed.). Berlin—New York: Mouton de Gruyter, 1992. P. 17—32.
35. *Sinclair J.* EAGLES Preliminary recommendations on Corpus Typology, 1996. Режим доступа: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html (дата обращения: 10.11.2020).
36. *Martí A.Mª.A., I. Castellón Masalles* Lingüística computacional. Barcelona: Edicions Universitat de Barcelona, 2000.
37. *Santalla del Río, M.ª P.* La elaboración de corpus lingüísticos // Nuevas tecnologías en Lingüística, Traducción y Enseñanza de lenguas / Cal M., Núñez P., Palacios I.M. (eds.). Santiago de Compostela: Universidade de Santiago de Compostela, Servizo de Publicacións e Intercambio Científico, 2005. P. 45—63.
38. *Захаров В.П.* Корпусная лингвистика. СПб., 2005.

39. *Козлова Н.В.* Лингвистические корпуса: определение основных понятий и типология // Вестник Новосибирского государственного университета. Серия: Лингвистика и коммуникация. 2013. № 11(1). С. 76—89.
40. *Кибрик А.Е., Брыкина М.М., Леонтьев А.П., Хитров А.Н.* Русские посессивные конструкции в свете корпусно-статистического исследования // Вопросы языкознания. 2006. Вып. 1. С. 16—45.
41. *Torruella J., Llisterri J.* Diseño de corpus textuales y orales // Filología e informática. Nuevas tecnologías en los estudios filológicos / J.M. Blecua, G. Clavería, C. Sánchez, J. Torruella (eds.). Barcelona: Milenio Universidad Autónoma de Barcelona, Dpto. de Filología Española, 1999. P. 45—77.
42. *Кривнова О.Ф.* Области применения речевых корпусов и опыт их разработки // Сборник трудов XVIII Сессии Российского акустического общества РАО. Таганрог, 2006. С. 81—84.
43. PRESEEA.Corpus del Proyecto para el estudio sociolingüístico del español de España y de América. Alcalá de Henares: Universidad de Alcalá, 2014. Режим доступа: http://preseea.linguas.net (дата обращения: 01.11.2020).
44. *Рыков В.В.* Корпус текстов как реализация объектно-ориентированной парадигмы // Труды международного семинара «Диалог-2002». М.: Наука, 2002. С. 124—129.

References

1. Melnikov, G.P. (2003). *System typology of languages: Principles, methods, models*. Moscow: Nauka. (In Russ.).
2. Plungyan, V.A. (2008). Corpus as a tool and as ideology: on some topics of modern corpus linguistics. *Russian language in scientific coverage*, 2(16), 7—20. (In Russ.).
3. Moure, T. & Llisterri, J. (1996). Lenguaje y nuevas tecnologías: el campo de la lingüística computacional In M. Fernández Pérez (coord.) *Avances en Lingüística aplicada, Universidade de Santiago de Compostela, Servicio de Publicacións e Intercambio Científico*. Santiago de Compostela: Universidade de Santiago de Compostela, Servicio de Publicacións e Intercambio Científico. pp. 147—227. (In Spanish).
4. Real Academia Española (2001). Diccionario de la lengua española. Madrid: Espasa. URL: https://dle.rae.es/corpus (accessed: 10.11.2020). (In Spanish).
5. Ushakov, D.N. (2012). *Explanatory dictionary of the Russian language*. URL: https://gufo.me/dict/ushakov/корпус (accessed: 29.10.2020).
6. Villayandre Llamazares, M. (2008). Lingüística con corpus. *Estudios humanísticos. Filología*, 30, 329—349.
7. McEnery, T. (2003). Corpus Linguistics In en R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics.* Oxford: Oxford University Press. pp. 448—463.
8. McEnery, T. & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
9. McEnery, T. & Wilson, A. (2001). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
10. McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-Based Language Studies*. An advanced resource book, London-New York: Routledge. URL: https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/chapters/B03.pdf (accessed: 07.11.2020).
11. Chomsky, N. (1969). Quine's empirical assumptions. In D. Davidson & J. Hintikka (Eds.) *Words and objections. Essay on the Work of W.V Quine.* Dordrecth: D. Reidel. pp. 53—68.
12. Chomsky, N. (2006). *Language and mind*. Cambridge.
13. Abercrombie, D. (1965). *Studies in Phonetics and Linguistics*. London: Oxford University Press. URL: http://www.davidcrystal.com/Files/BooksAndArticles/-4896.pdf (accessed: 06.11.2020).
14. Juilland, A.G., Brodin, D.R. & Davidovitch, C. (1970). *Frequency dictionary of French words*. Hague—Paris: Mouton.
15. Biber, D., Conrad, S., Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
16. Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic computing*, 8(4), 243—257.

17. Baker, P., Hardie, A. & McEnery, T. (2006). *Glossary of Corpus Linguistics*. Edinburgh: University Press.
18. Krasina, E.A. & Novikova, M.L. (2019). Phenomenon of language in the paradigms of functional semantics and linguosemiotics (V Novikov readings. Moscow, April 18—19, 2019). *Russian Journal of Linguistics,* 23(3), 856—864. DOI: 10.22363 / 2312-9182-2019-23-3-856-864. (In Russ.).
19. Svartvik, J. (1992). Corpus linguistics comes of age In J. Svartvik (ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82 (Stockholm, 4—8 August, 1991).* Berlin—New York: Mouton de Gruyter. pp. 7—13.
20. Leech, G. (1992). Corpora and theories of linguistic performance In J. Svartvik (ed.) *Directions in Linguistics: Proceedings of Nobel Symposium 82 (Stockholm, 4—8 August, 1991).* Berlin—New: Mouton de Gruyter. pp. 105—122.
21. Labov, W. (1969). The logic of non-standard English. *Georgetown. Monographs on Language and Linguistics*, 22.
22. Krasina, E.A. & Perfilieva, N.V. (2018). Semantic parameters of quantitative units in different-structured languages. *Cognitive linguistics issues*, 1(54), 126—136. DOI: 10.20916/1812-3228-2018-1-126-136. (In Russ.).
23. Denisenko, V.N., Krasina, E.A. & Perfilieva, N.V. (2016). The principle of double meaning in language and word. *Cognitive linguistics issues,* 3(48), 103—108. (In Russ.).
24. Aarts, J. & Meijs, W. (eds.) (1984). *Corpus Linguistics*. Amsterdam: Rodopi.
25. Manual for the Corpus of Early English Correspondence Sampler CEECS (1998) Nurmi A. (ed.). Helsinki. URL: http://www.eng.helsinki.fi/doe/projects/ceec/ (accessed: 06.11.2020).
26. Taavitsainen, I. & Pahta, P. (1997). Corpus of Early English Medical Writing. *Computers in English Linguistics*, 21, 71—79.
27. Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford.
28. Stubbs, M. (2001). *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.
29. Stubbs, M. (2006). Corpus analysis: the state of the art and three types of unanswered question In Hunston, S. & Thompson, G. (eds.) *System and corpus: Exploring connections*. London: Equinox. pp. 15—36.
30. Simpson, R. & Swales, J. (2001). Introduction to North American perspective on corpus linguistics at the millennium In R. Simpson and J. Swales (eds.) *Corpus linguistics in North America. Selections from the 1999 Symposium*. Ann Arbor: The University of Michigan Press. pp. 1—14.
31. Parodi, G. (2008). Lingüística de Corpus: Una introducción al ámbito. *Revista de Lingüística Teórica y Aplicada*, 46(1), 93—119. (In Spanish).
32. Abaitua, J. (2002). Tratamiento de corpora bilingües In M.A. Martí & J. Llisterri (eds.) *Tratamiento del lenguaje natural: tecnología de la lengua oral y escrita*. Soria—Barcelona: Fundación Duques de Soria/Edicions de la Universitat de Barcelona. pp. 61—90. (In Spanish).
33. Aijmer, K. & Altenberg, B. (eds.) (1991). English Corpus Linguistics In *Studies in Honour of Jan Svartvik.* London: Longman.
34. Francis, W.N. (1992). Language Corpora B.C. In J. Svartvik (ed.) *Directions in Linguistics: Proceedings of Nobel Symposium 82* (Stockholm, 4—8 August 1991). Berlin—New York: Mouton de Gruyter. pp. 17—32.
35. Sinclair, J. (1996). *EAGLES Preliminary recommendations on Corpus Typology*. URL: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html (accessed: 01.11.2020).
36. Martí Antonín, Mª.A. & Castellón Masalles I. (2000). *Lingüística computacional*. Barcelona: Edicions Universitat de Barcelona. (In Spanish).
37. Santalla del Río, M.ª P. (2005). "La elaboración de corpus lingüísticos", en M. Cal, P. Núñez, I. M. Palacios (eds.): *Nuevas tecnologías en Lingüística, Traducción y Enseñanza de lenguas*, Universidade de Santiago de Compostela, Servizo de Publicacións e Intercambio Científico, 45—63. (In Spanish).
38. Zakharov, V.P. (2005). *Corpus linguistics*. Saint Petersburg. (In Russ.).

39. Kozlova, N.V. (2013). Linguistic corpus: definition of basic concepts and typology. *Novosibirsk State University Bulletin. Series: Linguistics and Communication*, 11(1), 76—89. (In Russ.).
40. Kibrik, A.E., Brykina, M.M., Leontiev, A.P. & Khitrov, A.N. (2006). Russian possessive constructions in the light of corpus-statistical research. *Questions of linguistics*, 1, 16—45. (In Russ.).
41. Torruella, J. & Llisterri, J. (1999). Diseño de corpus textuales y orales In J.M. Blecua, G. Clavería, C. Sánchez & J. Torruella (eds.) *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Milenio Universidad Autónoma de Barcelona, Dpto. de Filología Española. pp. 45—77. (In Spanish).
42. Krivnova, O.F. (2006). Areas of application of speech corpora and experience of their development In *Proceedings of the XVIII Session of the Russian Acoustic Society of RAO*. Taganrog. pp. 81—84. (In Russ.).
43. PRESEEA (2014). *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América.* Alcalá de Henares: Universidad de Alcalá. URL: http://preseea.linguas.net (accessed: 01.11.2020). (In Spanish).
44. Rykov, V.V. (2002). Text corpus as an implementation of the object-oriented paradigm In *Proceedings of the international seminar "Dialogue-2002".* Moscow: Nauka. pp. 124—129. (In Russ.).

Сведения об авторе:
Чилингарян Камо Павелович, старший преподаватель Института Гостиничного бизнеса и туризма Российского университета дружбы народов. *Сфера научных интересов*: сравнительная грамматика, корпусная лингвистика, иностранные языки, перевод; *e-mail*: chilingaryan-kp@rudn.ru; https://orcid.org/0000-0002-3863-8603; eLIBRARY SPIN-код 1650-3199

Information about the author:
Kamo P. Chilingaryan, Senior lecturer in Hotel business and tourism institute at RUDN University. *Research interests*: Comparative linguistics, corpus linguistics, foreign languages, translation/interpretation; *e-mail:* chilingaryan-kp@rudn.ru; https://orcid.org/0000-0002-3863-8603; eLIBRARY SPIN-code 1650-3199.