Научная статья
УДК 82`0 (5 Каз–Рус)
https://doi.org/10.31489/2024Ph2/96-105

Получено: 03.12.2023 г.
Принято: 04.04.2024 г.

Г.Ж. Кучумова*, Д.Р. Мухамеджанова, А.Б. Ибраева

Maqsut Narikbayev University, Астана, Казахстан
(E-mail: g_kuchumova@kazguu.kz)

Корпусная лингвистика как методология проведения эмпирических языковых исследований и ее развитие в Казахстане

В современной лингвистической науке широкое распространение приобретает методология корпусного анализа, с помощью которого возможно проведение эмпирических лингвистических исследований на базе корпусов или цифровых данных текстового материала. Настоящая статья посвящена раскрытию основных принципов и подходов корпусной лингвистики, а также анализу ее развития в казахстанской лингвистической науке. Для нахождения ответа на поставленную эмпирическую задачу нами был проведен контент-анализ статей, связанных с корпусной лингвистикой или корпусным анализом и опубликованных в двух казахстанских журналах. Результаты анализа показали, что использование методологии корпусного анализа слабо распространено среди казахстанских лингвистов, однако интерес к этой методологии начинает расти. Кроме того, было выявлено, что статьи, использующие корпусный анализ, лимитированы в применении разнообразных и более продвинутых методов корпусной лингвистики, что говорит о необходимости усовершенствования методологической подготовки лингвистов-корпусников в Казахстане. Авторами также предложены рекомендации по развитию корпусной лингвистики в стране.

Ключевые слова: корпусная лингвистика, корпусный анализ, корпус, лингвистическая наука, лингвист-корпусник.

Введение

В зарубежной лингвистической науке для проведения языковых исследований широко используется методология корпусной лингвистики, или раздела языкознания, занимающегося созданием и анализом текстовых корпусов [1; 2]. Становление корпусной лингвистики в современном ее понимании, то есть когда язык изучается с помощью компьютерных технологий, связывают с развитием компьютерных технологий [2; 16], благодаря которым стало возможно быстро создавать и хранить огромные базы данных текстового материала.

В зарубежной литературе корпусная лингвистика чаще всего понимается как методология проведения языковых исследований на основе анализа текстовых корпусов [1; 3]. Под корпусом понимается совокупность аутентичных текстов или образцов языкового пользования, представляющая собой компьютеризированную базу данных и используемая в целях лингвистических исследований [3; 21]. Методологические особенности корпусной лингвистики во многом определяют ее растущую популярность в языковой науке, что подтверждается появлением новых корпусов на различных языках. Первый электронный лингвистический корпус — *the BROWN corpus* — был создан в 1961 году в Брауновском университете (США) [4; 15]. Этот корпус, содержащий один миллион слов, состоит из текстов на американском варианте английского. В настоящее время существуют лингвистические корпуса, созданные как на основе других вариантов английского языка, так и на других языках, например, армянском, грузинском, испанском, казахском, немецком, португальском, румынском, русском, санскрите, турецком и т.д.

Сегодня корпусная лингвистика и ее аналитические методы используются для проведения эмпирических исследований в различных областях лингвистики. Настоящая статья раскрывает особенности корпусной лингвистики как методологии проведения эмпирических языковых исследований, а также обсуждает современное состояние ее развития в Казахстане.

Материал и методы

Для понимания современного состояния развития корпусной лингвистики в Казахстане было решено проанализировать количество и содержание научных статей, использующих методологию

* Автор-корреспондент. E-mail: *g_kuchumova@kazguu.kz*

корпусного анализа и публикуемых в научных журналах страны. На наш взгляд, такие данные могут служить показателем использования корпусного анализа в исследованиях казахстанских лингвистов и, соответственно, развития корпусной лингвистической науки в Казахстане.

Для сбора данных были выбраны два научных журнала, входящих в перечень журналов, рекомендуемых Комитетом по обеспечению качества в науке и высшем образовании Министерства науки и высшего образования Республики Казахстан, для публикации результатов исследования:

1. «Вестник Казахского национального университета. Серия филологическая» (далее — Вестник КазНУ).
2. «Вестник Евразийского национального университета им. Л. Гумилева. Серия филология» (далее — Вестник ЕНУ).

Поиск интересующих нас статей осуществлялся на Интернет-сайтах выбранных журналов с использованием ключевого слова «корпус» (также на английском языке «corpus») в заголовках и текстах статей. Для окончательной выборки статей и их анализа применялся контент анализ заголовков, аннотаций и разделов статей, описывающих методологию исследования.

Обзор литературы

Особенности корпусной лингвистики

Появление и развитие корпусной лингвистики связано с теоретическим сдвигом в понимании природы языка. Лингвисты-корпусники придерживаются мнения, что языковые правила и их изменения связаны с тем, как язык используется на практике, а точнее в коммуникативном процессе [2; 14]. Соответственно, чтобы понять, как устроен и функционирует определенный язык, необходимо изучить его использование в коммуникативном контексте. Другими словами, знания о языке как системе можно получить, изучая язык в действии. Корпусная лингвистика позволяет это делать на основе анализа огромных баз данных текстового материала, насчитывающего не только миллионы, но и миллиарды слов. Таким образом, изучение языка в корпусной лингвистике опирается не на интуицию исследователя, а на эмпирический материал — корпус.

Несмотря на то, что корпусная лингвистика имеет дело с текстами (письменными или устными), по сравнению с другими текстовыми анализами (например, стилистический анализ), она обладает своими особенностями. Во-первых, в корпусной лингвистике исследователь не изучает текст как целое, а фокусирует свое внимание на его фрагментах или лингвистических единицах разного уровня в зависимости от поставленной задачи, например, словах, словосочетаниях или предложениях. Во-вторых, тексты, входящие в корпус, не изучаются лингвистом-корпусником с целью понятия их содержания, а анализируются лишь повторяющиеся языковые феномены или образцы языковой коммуникации. В-третьих, корпусная лингвистика дает возможность получить представление о языке как системе, в то время как текстовый анализ как таковой позволяет изучить определенный образец языка в действии [2; 18], выраженный, к примеру, одним литературным произведением или коллекцией трудов одного автора. Кроме того, корпусная лингвистика позволяет применять как количественный, так и качественный подход к лингвистическому анализу. Другими словами, можно анализировать семантику слова или словосочетания в языковом контексте и в то же время выявлять взаимосвязь языка и демографических данных его носителей.

Корпус и его характеристики

Для более ясного понимания как применяется методология корпусной лингвистики на практике необходимо понимать, что есть корпус. Корпус — это совокупность аутентичных текстов или образцов языкового пользования, представляющая собой компьютеризированную базу данных и используемая в целях лингвистических исследований [3; 2]. При этом тексты, входящие в корпус, могут быть как письменными, так и устными. В последнем случае образцы устной речи записываются на аудио, транскрибируются и собираются в корпус.

Лингвистический корпус должен в идеале обладать тремя характеристиками — быть аутентичным, репрезентативным и аннотированным. Под *аутентичностью* корпуса понимается его «естественность» или то, что коллекция текстов, входящая в корпус, должна быть представлена образцами «естественного» языка [1; 30] или языка, используемого в реальной коммуникативной ситуации. Однако аутентичность лингвистического корпуса не так легко обеспечить. Этот вопрос особо актуален при создании корпусов устных текстов, которые при транскрибировании теряют часть своих как лингвистических, так и паралингвистических особенностей, например, информацию о тональности речи или акценте говорящего. В случае, когда необходимо записывать устную речь для последующе-

го анализа, аутентичность может полностью теряться, так как под наблюдением участники исследования могут использовать менее «естественный» язык, к примеру, быть более лингвистически «аккуратными» при выражении своих мыслей или в общении друг с другом. Этот феномен Stefanowitsch [3; 25] называет *парадоксом наблюдателя* и советует при сборе лингвистических данных быть более креативным для обеспечения минимального нарушения «естественной» коммуникации участников исследования.

Под *репрезентативностью* корпуса понимается его «сбалансированность» [5; 13], а именно его способность отражать, например, определенный язык посредством его разнообразных видов коммуникативной вариативности или функциональных стилей, например, художественного, научного, публицистического, разговорного и т.д. Предполагается, что результаты исследования, основанного на репрезентативном корпусе определенного языка, могут быть обобщены на этот язык как систему. В этом отношении методология корпусной лингвистики схожа с количественной методологией социологических исследований. Соответственно, при создании лингвистического корпуса, исследователь должен обдуманно подходить к выборке текстов и структуре корпуса, если он планирует обобщить результаты своего анализа на определенный язык или функциональный стиль.

Кроме включения в корпус различных функциональных стилей или видов текста, репрезентативность корпуса обеспечивается его размером. Считается, что, чем больше корпус, тем более он сбалансирован. В идеале, чтобы корпус репрезентировал определенный язык или функциональный стиль, он должен соответствовать двум требованиям — обладать вариативностью языковых дискурсов или видов текста и быть достаточно большого размера [6; 79]. Например, Современный корпус американского английского (Corpus of Contemporary American English) состоит из восьми разных видов дискурса (тексты фильмов, блогов, интернета, устной речи, фантастики, журналов, газеты и научной литературы) и насчитывает один миллиард слов.

Для проведения комплексных и сложных анализов, корпус также должен быть *аннотирован*. Другими словами, его надо *реконтекстуализировать,* или подвергнуть лингвистической разметке [3; 38; 1; 32], то есть включить в корпус метаданные или информацию о паралингвистических свойствах текстов, входящих в корпус, лингвистических параметров слов, составляющих корпус, а также данные о создателях текстов или условиях, при которых тексты были созданы. К паралингвистическим свойствам письменных текстов относятся размер, цвет и стиль шрифта, специальные обозначения или символы. Для устных текстов обычно предоставляется информация об интонации, совмещении речи, длины пауз и т.д. Под лингвистическими параметрами слов понимается информация о их части речи и словарной форме. Если в корпус включена такая дополнительная информация, то корпус считается *таггированным* и *леммативированным* [7; 38]. Данные о создателях текстов обычно включают такую информацию, как возраст, пол, этническую принадлежность, образование, локацию и т.д. Данные об условиях создания текста могут включать жанр текста или описание ситуации, при которой текст был создан, например на шумной улице или в конференц-зале. Таким образом, можно сказать, что реконтекстуализация корпуса способствует его аутентичности, то есть дополнительная информация о текстах корпуса помогает лингвисту-корпуснику лучше понимать с каким корпусом он имеет дело.

Следует отметить, что аннотирование корпуса очень трудоемкий и долгий процесс [4; 6]. Однако сейчас уже есть пакеты программного обеспечения для анализа языковых корпусов, которые могут автоматически за пару минут таггировать и лемматизировать разработанный корпус. К такому программному обеспечению относится *LancsBox*, инновационный корпусный инструмент, разработанный в Университете Ланкастер (Великобритания) группой исследователей под руководством Brezina [7].

Типы корпусов

Корпуса создаются для проведения лингвистических исследований разнообразных по своей тематике и методологии. В зависимости от поставленных целей и задач исследования создаваемые корпуса могут быть различных типов. Так, по форме языка различают корпуса письменной речи и устной речи (тексты разговоров, бесед, дискуссий и т.п.) [8; 39, 9; 54]. По содержанию или тематике, корпуса бывают общие, то есть содержат тексты из разных тематических областей, например, как Британский национальный корпус (BNC: British National Corpus), или специализированные, то есть включающие группу текстов определенного тематического или профессионального направления, например, как Корпус мнений Верховного Суда США (Corpus of US Supreme Court Opinions) или Корпус фильмов (The Movie Corpus).

По языковому принципу корпуса бывают одноязычные, параллельные и сравнительные [1; 69]. Большинство корпусов являются одноязычными, то есть содержат текстовый материал на одном языке. Параллельные корпуса, напротив, состоят из двух и более одноязычных корпусов, при этом эти корпуса являются переводом друг друга. Другими словами, в параллельном корпусе каждый текст одного одноязычного корпуса имеет свой перевод в другом одноязычном корпусе, входящем в этот же параллельный корпус. Часто такие корпуса используются в переводческой практике. Примером такого корпуса служит Англо-норвежский параллельный корпус (ENPC: English-Norwegian Parallel Corpus). Сравнительные корпуса — это корпуса, состоящие также из двух и более одноязычных корпусов, которые содержат разные тексты, но разработаны по одному принципу, что позволяет сравнивать корпуса и их данные, так как они схожи в плане выборки текстов. Например, *the CHILDES corpus*, представляющий собой базу транскриптов детской речи на 24 языках, является сравнительным (CHILDES — child language corpus in many languages | Sketch Engine).

По временному принципу корпуса могут быть диахроническими, мониторинговыми и синхронными. Диахронические корпуса — это корпуса, которые создаются на базе текстов, относящихся к определенному историческому промежутку. Например, Корпус исторического американского английского (COHA: corpus of Historical American English) содержит сто тысяч текстов (или 400 миллионов слов), опубликованных в период с 1810 по 2009 год (English Corpora: most widely used online corpora. Billions of words of data: free online access (english-corpora.org)). Этот корпус позволяет лингвистам исследовать широкий спектр исторических изменений в американском английском с большой точностью. Временной подход к разработке корпуса также присущ для мониторинговых корпусов. Главное отличие их только в том, что такие корпуса регулярно или постоянно пополняются и обновляются. Другими словами, такие корпуса «растут». Примером такого корпуса является Корпус интернет-новостей (NOW: News on the Web Corpus). Работа над созданием этого корпуса началась в 2010 году. Сейчас он насчитывает более 15 миллиардов слов данных, собранных из 21 англоязычной страны. Сам корпус растет примерно на 200–220 миллионов слов в месяц (English Corpora: most widely used online corpora. Billions of words of data: free online access (english-corpora.org)). Синхронные корпуса создаются на основе текстового материала, принадлежащего к одному историческому времени. Например, ранее упомянутый *the BROWN corpus* состоит из текстов, опубликованных в 1961 году (BROWN Corpus search online | Sketch Engine).

Следует отметить, что нелегко отнести корпус к определенному типу. Один и тот же корпус может одновременно относиться к разным типам.

Применение корпусного анализа

Корпусный анализ применяется практически во всех направлениях современной лингвистики. В лингвистической науке его первыми пользователями были лексикографы [4; 3], изучающие семантику слов и фраз и их использование в речи для составления лингвистических словарей. Сейчас же корпусный анализ используется в вопросах исторического развития языка, изучении диалектов, анализе особенностей стилей речи, грамматике, фонетике и фонологии, фразеологии, социолингвистике, изучении политического дискурса и т.д. Примерами таких исследований являются: (1) статья Jiang и Hyland [10], в которой авторы представили результаты анализа меняющейся обеспокоенности международной прессы в течение 2020 года о разворачивающихся событиях, связанных с пандемией COVID-19; (2) статья Zawadzka-Paluektau [11], рассматривающая репрезентацию украинских беженцев в польской прессе в начале российского вторжения 2022 года; (3) статья Yến-Khanh [12], в которой исследователь рассказывает как средства массовой информации Вьетнама освещают проблемы аутизма, являющегося общественной проблемой страны; (4) статья Willis [13], где представлены результаты исследования о том, как британские политики обсуждают вопросы, связанные с изменением климата; (5) статья Tagliamonte и Smith [14] о том, как использование наречия *obviously* меняется во времени на территории Великобритании и Канады; (6) статья Adam [15], раскрывающая взгляды университетов Великобритании на внедрение технологий в преподавание и обучение.

Корпусный анализ применяется также в практической деятельности. Например, в области преподавания языков, корпусный анализ используется для разработки учебного материала или выявления неких стандартов применения языка в коммуникации для составления контрольных заданий. Корпусный анализ также используется в переводческой деятельности для повышения качества и скорости перевода. В судебной экспертизе корпусный анализ применяется для подтверждения авторства [4; 7].

Основные виды корпусного анализа

Методы корпусного анализа различны и могут базироваться как на количественном, так и на качественном подходе к исследуемому материалу. В настоящей статье мы описываем основные методы корпусного анализа, а именно частотный анализ, дистрибутивный анализ, анализ коллокаций и конкордансы.

Под *частотным анализом* в корпусной лингвистке подразумевается подсчет частоты использования определенного слова-токена или словоформы в корпусе. Если корпус лемматизирован, то частотный анализ также показывает частоту использования определенной леммы [7; 42]. Зачастую корпусный анализ начинается с метода частотного анализа, потому что он показывает описательную статистику о корпусе, а именно, какие слова входят в корпус и в каком количестве, а также какие слова наиболее или наименее частотны.

Частота использования токена или леммы в корпусе может выражаться в двух статистических мерах — абсолютной и относительной [3; 143]. Абсолютная частота — это количество употреблений словоформ в корпусе. Такая частота выражается целым числом. Относительная частота показывает отношение количества употреблений определенной словоформы к количеству всех словоформ в корпусе, то есть его размеру. Относительная частота, также известная как *нормализованная*, является дробным числом. Например, в *the BROWN corpus* абсолютная частота артикля *the* составляет 69 971, или этот артикль встречается в этом корпусе 69 971 раз. Его же относительная частота равна числу 59 515,6. Данное число было получено при делении абсолютной частоты артикля *the* на размер корпуса, то есть количество всех его словоформ, а именно 1 007 299 токенов, и умножении полученного числа на взятую основу *нормализации*, в данном случае число 1 000 000, так как корпус состоит из более одного миллиона токенов. Следует отметить, что нормализация — это способ корректировки подсчета абсолютной частоты токена в корпусах разного размера, чтобы их можно было точно сравнивать [16; 263].

Применение частотного анализа в исследовательской практике можно продемонстрировать исследованием, выполненным Alexioul и Kokla [17], которые изучили, насколько мультсериал «Свинка Пеппа» подходит для изучения английского языка для начинающих. Они сравнили словарный запас мультфильма (корпус 1) со списком слов программы, рассчитанной для детей начальной и средней школы, а именно готовящихся к сдаче Cambridge Young Learners English Test (корпус 2). Сравнивая слова и их абсолютную частоту в двух корпусах, они пришли к выводу, что просмотр мультсериала может способствовать овладению уровнем Cambridge English: Young Learners, так как мультсериал использует 84,87 % словарного запаса из списка слов, необходимого для сдачи Cambridge Young Learners English Test.

Дистрибутивный анализ показывает распределение токена или леммы в корпусе [18; 122], а точнее по его текстам или категориям текстов, организованным по какому-либо признаку, например, стилю речи, году, гендерному признаку, локации и т.д. Другими словами, дистрибутивный анализ показывает, насколько равномерно словоформы распределены в текстах корпуса. Этот тип анализа также известен под названием дисперсионный анализ. Стандартное отклонение (*sd*) является классической мерой дисперсии [7; 10]. Например, две разные словарные формы *a* и *b* могут использоваться в одном и том же корпусе с одинаковой частотой. Однако частота применения *a* в каждом тексте или категории текстов корпуса может быть одинаковой, в то время как *b* может встречаться только в одном тексте или категории текстов корпуса. Другими словами, можно утверждать, что по сравнению со словоформой *b*, словоформа *a* равномерно распределена в корпусе. Дистрибутивный анализ широко применяется в создании словарей или словарных списков на базе корпуса [19; 116]. Рисунок 1 демонстрирует результаты дистрибутивного анализа словоформы *news* в Корпусе Интернет-новостей (NOW: News on the Web Corpus)

Frequency by country (Return to frequency by year)

SECTION	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
United States	3311447	6,804.3	486.67	
Canada	944182	2,018.3	467.82	
Great Britain	991457	2,337.7	424.11	
Ireland	706682	1,184.4	596.65	
Australia	590094	1,302.3	453.11	
New Zealand	332819	645.7	515.48	
India	1004574	1,859.9	540.14	
Sri Lanka	49751	136.1	365.59	
Pakistan	258624	386.8	668.67	
Bangladesh	49664	97.0	512.14	
Malaysia	270249	378.2	714.48	
Singapore	452312	610.7	740.65	
Philippines	319796	490.4	652.15	
Hong Kong	104326	83.9	1,243.72	
South Africa	689588	796.6	865.72	
Nigeria	606228	887.9	682.75	

Рисунок 1. Результаты дистрибутивного анализа словоформы *news* по географическому признаку на базе Корпуса Интернет-новостей (NOW: News on the Web Corpus)

Анализ коллокаций, или словосочетаний двух и более слов, имеющих признаки синтаксически и семантически целостной единицы, применяется для выявления слов, часто используемых друг с другом в корпусе [20; 107], что применимо, например, в преподавании или изучении иностранного языка или в переводческой деятельности. Кроме того, анализ коллокаций способствует пониманию *атмосферы* лингвистического контекста, в которой используется слово [21]. Другими словами, изучая с какими словами коллоцируется определенная словоформа, можно понять, как феномен или процесс, обозначаемый этой словоформой, репрезентируется, например, в политическом дискурсе или в языке газеты. В корпусной лингвистке используется несколько статистических мер коллокаций. Некоторые из них, такие как логарифмическая вероятность *(LL)* и *t*-показатель (*t-score*) основываются на частоте коллокации в корпусе. Другие, например, *z-показатель (z-score)* и *MI-показатель (MI-score)* базируются на эксклюзивности коллокации (рис. 2)

Рисунок 2. Меры анализа коллокаций [7; 74]

Примером использования анализа коллокаций может служить исследование Willis [13], в котором автор изучал политический дискурс британских политиков, и как в нем представлена проблема изменения климата. Так, например, анализируя коллокации словоформы *carbon*, было выявлено, что она чаще всего используется с такими словами, как *storage, capture, low, budget, price* и *economy.* Автор пришел к выводу, что проблема изменения климата обсуждается в британском политическом дискурсе в техническом плане, то есть ссылаясь на исследования, описывая ее природу, говоря о ее последствиях на окружающую среду и экономику, в то время как социальный компонент проблемы, а

именно влияние изменения климата на здоровье и жизнедеятельность людей, не получает широкого обсуждения.

Конкордансы — это отрезки текстов, которые формируются на основе поискового слова или фразы [5; 32]. Анализ конкордансов позволяет собрать большое количество примеров использования определенной словоформы в ее исходном контексте в одном месте. Рисунок 3 демонстрирует список конкордансов со словоформой *eye*.

Анализ конкордансов, по сути, основан на качественном подходе к исследованию. Другими словами, исследователь их рассматривает (читает), классифицирует по интересующему его принципу, обобщает примеры похожих употреблений интересуемой его словоформы и интерпретирует их, представляя свои результаты в качестве определенных тематических категорий.

Рисунок 3. Список конкордансов со словоформой *eye* на примере Британского национального корпуса (BNC: British National Corpus)

Результаты и обсуждение

Развитие корпусной лингвистики в Казахстане

Поиск статей о корпусной лингвистике, или использующих методологию корпусного анализа, на Интернет-сайтах выбранных журналов показал 32 результата, которые включают 23 статьи Вестника Казахского национального университета и 9 статей Вестника Евразийского национального университета. После изучения заголовков, аннотаций и методологии этих статей 9 статей были исключены из дальнейшего анализа, так как они не были связаны с корпусной лингвистикой или корпусным анализом. Термин «корпус» в этих статьях использовался для обозначения некой базы текстов, собранных авторами для исследования вручную, или же для отсылки к словарному запасу анализируемого языка. Таким образом, только 23 статьи были выбраны для более детального дальнейшего контент-анализа.

Результаты анализа показали, что выбранные статьи были опубликованы в разные года. В Вестнике КазНУ статьи, относящиеся к корпусной лингвистике и корпусному анализу, вышли в свет в период с 2015 по 2019 год. Наибольшее количество из них было опубликовано в 2015 и 2016 годах. В Вестнике ЕНУ статьи о корпусной лингвистике, или использующие корпусный анализ, начались публиковаться относительно недавно с 2021 года. Согласно данным, их количество растет. По сравнению с 2022 годом, в 2023 году было выпущено 3 статьи только в первых двух томах журнала. Детальная информация о количестве и периоде статей выборки представлена в таблицах 1 и 2.

Таблица 1

Количество и период публикации статей, относящихся в корпусной лингвистике, в Вестнике КазНУ

2015 г.	2016 г.	2017 г.	2018 г.	2019 г.
5	7	0	2	1

Таблица 2

Количество и период публикации статей, относящихся в корпусной лингвистике, в Вестнике ЕНУ

2021 г.	2022 г.	2023 г.
1	4	3

Однако при сравнении количества статей, относящихся к корпусной лингвистике, с количеством статей, опубликованных в этот же период в целом, получается, что их количество значительно меньше. Так, в Вестнике КазНУ доля таких статей составляет лишь 1,3 %, то есть 15 из 1 211. В Вестнике ЕНУ доля интересуемых нас статей составляет около 5 %, точнее 8 из 163. Это говорит о том, что в Казахстане методология корпусного анализа не используется широко. Интерес к ней только начинает расти.

Анализ заголовков и аннотаций статей показал, что статьи написаны по разным направлениям лингвистики, а именно по социолингвистике, теории перевода, семантике, лексикологии, лексикографии, прагмалингвистике и методике обучения иностранному языку. Кроме того, согласно полученным результатам, большинство статей выборки (15 из 23) носят эмпирический характер, то есть основываются на анализе имеющихся корпусов (например, Национальном корпусе русского языка, Британском национальном корпусе, Современном корпусе американского английского) или самостоятельно созданных корпусов. На наш взгляд, это свидетельствует о том, что в Казахстане есть лингвисты, хотя и небольшое количество, которые обладают знаниями и навыками работы с корпусами, а также могут создавать свои собственные корпуса.

Однако детальный анализ описания методологии, использованной в рассматриваемых статьях, показал, что большинство авторов статей очень кратко описывают, как они проводили корпусный анализ, а именно недостаточное внимание уделяют описанию используемого корпуса (подкорпуса) и примененных методов корпусного анализа. Необходимо отметить, что краткое описание методологии исследования приводит к непониманию как применялся тот или иной метод корпусного анализа, что негативно сказывается на валидности и надежности самого исследования. Более того, эмпирические статьи в основном опираются на частотный анализ и анализ конкордансов, что, возможно, говорит о недостаточной методологической подготовке казахстанских лингвистов применять разнообразные и более сложные методы корпусного анализа с применением статистических мер. Последнее, необходимо помнить, что корпусный анализ позволяет предоставлять количественные данные не только касательно частотного анализа, но и конкордансов, коллокаций и других видов анализа. Однако в некоторых статьях такие количественные данные отражаются не в полной мере, что также ограничивает понимание всего процесса исследования, например выборку коллокаций или конкордансов.

Заключение

В настоящей статье были рассмотрены основы корпусного анализа и вопросы развития корпусной лингвистики в Казахстане. Важность корпусного анализа для проведения эмпирических лингвистических исследований растет. Интерес к нему также начинают проявлять и лингвисты Казахстана, которые применяют корпусный подход в своих исследованиях. Однако, как показали результаты анализа двух ведущих казахстанских журналов в области филологии, методологическая подготовка по корпусному анализу и написанию статей, основанных на методологии корпусной лингвистики, требует дальнейшего развития и усовершенствования.

Навыки проведения корпусного анализа могут быть развиты в рамках вузов. Для этого мы рекомендуем, во-первых, систематично проводить семинары или мастер-классы по корпусной лингвистике для исследователей университетов. На наш взгляд, для повышения качества корпусного анализа необходимо повышать знания лингвистов в области статистики, а также навыков использования в исследовании специальных программ для корпусного анализа, например *LancsBox* и *AntConc*. Во-вторых, для развития лингвистов-корпусников в стране необходимо внедрять продвинутые курсы корпусной лингвистики на уровне магистратуры и докторантуры. В-третьих, для повышения уровня владения методологией корпусной лингвистики нашим лингвистам необходим доступ к зарубежным журналам, публикующих статьи по корпусной лингвистике. Например, следующие журналы имеют высокую репутацию в мировом научном сообществе, поэтому доступ к ним может послужить некой площадкой изучения международного опыта применения корпусного анализа в прикладных лингвистических исследованиях:

1. Applied Linguistics (Q1).
2. International Journal of Corpus Linguistics (Q1).
3. International Journal of Applied Linguistics (Q1).
4. Corpus Linguistics and Linguistic Theory (Q1).
5. Applied Linguistics Review (Q1).
6. Australian Review of Applied Linguistics (Q1).
7. European Journal of Applied Linguistics (Q1)).

В-четвертых, необходимо налаживать международные контакты с лингвистами-корпусниками для обмена опытом и проведения совместных исследований, что может также положительно отразиться на повышении профессиональной компетенции казахстанских лингвистов-корпусников. Последнее, вузы могут организовать курсы повышения квалификации для них зарубежом. Например, ежегодная Международная летняя школа, проводимая по корпусной лингвистике в Университете Ланкастер (Великобритания) для начинающих исследователей, как раз и служит для таких целей.

Список литературы

1 McEnery, T., & Wilson, A. (2001). Corpus Linguistics: An Introduction. Edinburg: Edinburg University Press.

2 Bonelli, E.T. (2010). Theoretical overview of the evolution of corpus linguistics. In M. McCarthy & A. O'Keeffe (Eds.). *The Routledge Handbook of Corpus Linguistics* (pp. 14–28). NY: Routledge.

3 Stefanowitsch, A. (2020). *Corpus Linguistics: A Guide to the Methodology*. Berlin: Language Science Press.

4 McCarthy, M., & O'Keeffe, A. (2010). Historical perspective: What are corpora and how have they evolved? In M. McCarthy & A. O'Keeffe, (Eds.). *The Routledge Handbook of Corpus Linguistics* (pp. 3–13). NY: Routledge.

5 Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

6 Crawford, W., & Csomay, E. (2016). *Doing Corpus Linguistics.* NY: Routledge.

7 Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.

8 Adolphs, S., & Knight, D. (2010). Building a spoken corpus: What are the basics? In M. McCarthy & A. O'Keeffe (Eds.). *The Routledge Handbook of Corpus Linguistics* (38–52). Routledge.

9 Nelson, M. (2010). Building a written corpus: What are the basics? In M. McCarthy & A. O'Keeffe (Eds.). *The Routledge Handbook of Corpus Linguistics* (pp. 53-65). NY: Routledge.

10 Jiang, F.K., & Hyland, K. (2022). COVID-19 in the news: The first 12 months. *International Journal of Applied Linguistics, 32*(2), 241–258 https://doi.org/10.1111/ijal.12412

11 Zawadzka-Paluektau, N. (2022). Ukrainian refugees in Polish press, *Discourse and Communication*, *17*(1), 96–111. https://doi.org/10.1177/17504813221111636

12 Yến-Khanh, N. (2023). Representation of autism in Vietnamese digital news media: A computational corpus and framing analysis. *Communication Research and Practice*, *9*(2), 142–158. https://doi.org/10.1080/22041451.2023.2167510

13 Willis, R. (2017). Taming the climate? Corpus analysis of politicians' speech on climate change. *Environmental Politics, 26*(2), 212–231 https://doi.org/10.1080/09644016.2016.1274504

14 Tagliamonte, S., & Smith, J. (2021). *Obviously* undergoing change: Adverbs of evidentiality across time and space. *Language Variation and Change, 33*(1), 81–105. https://doi.org/10.107/S0954394520000216

15 Adam, M. (2021). Sociotechnical imaginaries in the present and future university: A corpus-assisted discourse analysis of UK higher education texts. *Learning, Media and Technology*, *46*(2), 204–217. https://doi.org/10.1080/17439884.2021.1864398

16 Biber, D., Conrad, S., & Reppen, R. (1998). Norming frequency counts. In *Corpus Linguistics: Investigating Language Structure and Use* (pp. 263–264). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511804489.017

17 Alexioul, T., & Kokla, N. (2019). Cartoons that make a difference: A linguistic analysis of Peppa Pig. *Journal of Linguistics and Education Research, 1*(1), 24–30. https://doi.org/10.30564/jler.v1i1.314

18 Egbert, J., Biber, D., & Gray, B., (2022). Designing and evaluating language corpora: A practical framework for corpus representativeness. Cambridge: Cambridge University Press.

19 Gries, S.T. (2020). Analyzing dispersion. In M. Paquot & S.T. Gries (Eds.). *A Practical Handbook of Corpus Linguistics* (pp. 99–118). Springer. https://doi.org/10.1007/978-3-030-46216-1_5

20 Xiao, Z. (2015). Collocation. In D. Biber & R. Reppen, (Eds.). *The Cambridge Handbook of Corpus Linguistics* (pp. 106–124). Cambridge: Cambridge University Press.

21 Baker, P., Gabrielatos, C., & McEnery, T. (2013). Sketching Muslims: A corpus driven analysis of representations around the word 'Muslim' in the British press 1998-2009. *Applied Linguistics, 34*(3), 255–278. https://doi.org/10.1093/applin/ams048.

Г.Ж. Кучумова, Д.Р. Мухамеджанова, А.Б. Ибраева

Корпустық лингвистика эмпирикалық тілдік зерттеулер жүргізу әдістемесі ретінде және оның Қазақстанда дамуы

Қазіргі лингвистикалық ғылымда корпустық талдау әдістемесі кең таралуда. Оның көмегімен корпустық немесе цифрлық мәтіндік материалдың деректеріне негізделген эмпирикалық лингвистикалық зерттеулер жүргізуге болады. Мақала корпустық лингвистиканың негізгі принциптері мен тәсілдерін ашуға, сондай-ақ оның қазақстандық лингвистика ғылымындағы дамуын талдауға арналған. Қойылған мақсаттарымызға жауап табу үшін біз екі қазақстандық журналда жарияланған корпустық лингвистикаға немесе корпустық талдауға қатысты мақалаларға контент-талдау жүргіздік. Талдау нәтижелері тіл мамандары арасында корпустық талдау әдістемесін қолдану кең таралмағанын, бірақ бұл әдістемеге қызығушылықтың арта бастағанын көрсетті. Сонымен қатар, корпустық талдауды қолданатын мақалалардың корпустық лингвистиканың әртүрлі және неғұрлым жетілдірілген әдістерін қолдануда шектеулі екені анықталды, бұл Қазақстандағы корпус лингвистерінің әдістемелік дайындығын жетілдіру қажеттігін көрсетеді. Мақалада елімізде корпус лингвистикасын дамытуға қатысты ұсыныстар да берілген.

Кілт сөздер: корпустық лингвистика, корпустық талдау, корпус, лингвистика, корпус лингвисі.

G. Kuchumova, D. Muhamejanova, A. Ibrayeva

Corpus linguistics as a methodology for conducting empirical language studies and its development in Kazakhstan

The methodology of corpus analysis is widely spreading today in modern linguistics. This methodology enables linguists to conduct empirical language studies based on corpora or computerized text data. In this regard, this paper presents foundational principles and approaches of corpus linguistics, as well as examines its development in Kazakhstan. The empirical part of the study was based on content analysis of articles related to corpus linguistics or corpus analysis and published in two Kazakhstani research journals. The analysis showed that local researchers seldom use corpus analysis in their studies; however, the interest in this methodology seems to be growing. Moreover, we found that the application of corpus analysis is limited in variety and to less sophisticated methods which suggests that there is a need for enhancing methodological training in corpus linguistics in Kazakhstan. Some recommendations are also suggested in the paper on how to increase research capacity in corpus linguistics in the country.

Keywords: corpus linguistics, corpus analysis, corpus, linguistics, corpus-linguist.

Information about authors

Kuchumova, Gulfiya Zhasulanovna — PhD, Associate Professor, School of Liberal Arts, Maqsut Narikbayev University, 010010, Astana, Kazakhstan, g_kuchumova@kazguu.kz;

Mukhamejanova, Dinara Ramazanovna — PhD, Associate Professor, School of Liberal Arts, Maqsut Narikbayev University, 010010, Astana, Kazakhstan, dinara_mukhamejanova@kazguu.kz;

Ibrayeva, Anar Baurzhanovna — PhD, Teaching Professor, School of Liberal Arts, Maqsut Narikbayev University, 010010, Astana, Kazakhstan, a_ibrayeva@kazguu.kz.