УДК 811.521

ИСТОРИЯ И СОВРЕМЕННОЕ СОСТОЯНИЕ КОРПУСНОЙ ЛИНГВИСТИКИ В ЯПОНИИ

Ю.В. Жданова, *Пермский федеральный исследовательский центр УрО РАН, Пермский государственный национальный исследовательский университет*

Для цитирования:
Жданова Ю.В. История и современное состояние корпусной лингвистики в Японии // Вестник Пермского федерального исследовательского центра. – 2025. – № 2. – С. 19–31. https://doi.org/10.7242/2658-705X/2025.2.2

Бурное развитие корпусной лингвистики в 1990-х годах кардинально изменило условия языковых исследований, предоставив европейским лингвистам новые возможности работы с обширными массивами языковых данных. В японской лингвистике увлекательный процесс создания национального корпуса языка сталкивался со значительными трудностями по многим причинам. Однако в русскоязычной лингвистической литературе на данный момент информации о состоянии корпусов японского языка явно недостаточно. Частично эту лакуну заполнила статья И.Л. Корецкой (2022), в которой сделана попытка описания основных корпусов японского языка на сегодняшний день. Тем не менее, история развития корпусов японского языка и его современное состояние требуют более тщательного описания и изучения по причине острой актуальности решения внутриязыковых проблем японского языка. В данной статье прослеживаются аспекты истории и развития корпусов японского языка, а также освещаются некоторые трудности, которые приходится преодолевать японским лингвистам на разных этапах создания японского национального корпуса языка.

Ключевые слова: *корпус японского языка, становление корпусной лингвистики, японская корпусная лингвистика, корпусы текстов, корпус письменного японского языка, корпус спонтанной речи, международный учебный корпус японского языка как иностранного.*

Начало развития корпусной лингвистики

С 1990-х годов корпусная лингвистика стала революционным направлением в изучении языков, предоставив исследователям возможность работать с обширными массивами реальных языковых данных.

Лингвистическим (языковым) корпусом называют крупный, машиночитаемый, унифицированный и структурированный массив языковых данных, размеченный и подготовленный с филологической точностью для решения конкретных лингвистических задач [1, 7].

Другое определение корпуса приведено на официальном сайте Национального корпуса русского языка (НКРЯ) (по состоянию на 31.03.2019 г.): «Корпус – это информационно-справочная система, основанная на собрании текстов на некотором языке в электронной форме.

Национальный корпус представляет данный язык на определенном этапе (этапах) его существования и во всем многообразии жанров, стилей, территориальных и социальных вариантов и т.п.» (http://www.ruscorpora.ru/corpora-intro.html).

Созданные за последние десятилетия многочисленные частотные словари и корпусы на материале различных языков, существенно отличающиеся от первых частотных списков немецких слов Ф.В. Кэдинга и П. Менцерата, в качественном и количественном отношении представляют интерес для множества лингвистических дисциплин и, благодаря значительному объему выборки, являются достоверным источником сведений о языке. Верхние страты частотных словарей, созданных на основе корпусов, можно считать отражением того «инварианта», которым владеют все носители языка, ядра языка, которым следует руководствоваться при изучении словарного состава, при определении существенных, ключевых, строевых для данного языка характеристик.

Корпусный подход упрочился в лингвистике последних десятилетий в силу целого ряда преимуществ: репрезентативный объем корпусов гарантирует типичность данных и способствует полному представлению всего спектра языковых явлений; кроме того, данные разного типа приводятся в корпусе в контекстной форме, что создает возможность их всестороннего и объективного изучения; причем массив данных может использоваться многократно, многими исследователями для решения разных задач [1, 7].

Сегодня существует множество национальных и репрезентативных корпусов языков мира: Национальный корпус русского языка, Британский национальный корпус, Американский национальный корпус, Мангеймский корпус немецкого языка, Корпус немецкого языка, Корпус французского языка, Венгерский национальный корпус, Корпус современного китайского языка и др.

Национальные корпусы, как правило, включают не менее 100 млн. словоупотреблений, охватывая широкий спектр жанров, стилей, а также региональных и социальных разновидностей языка. Весь материал в корпусе упорядочен: зафиксированы позиции слов в предложениях и частотность их употребления [2, 102–103].

Корпусные исследования обладают рядом преимуществ. Во-первых, большой объём данных обеспечивает репрезентативность и полноту описания языковых явлений. Во-вторых, данные представлены в естественном контексте, что способствует объективному анализу. В-третьих, однажды созданный корпус может многократно использоваться разными исследователями в различных целях [1, 3].

История развития языковых корпусов включает несколько ключевых этапов. История корпусной лингвистики берёт начало в середине XIX века, когда немецкий лингвист Эдуард Форстеманн в 1852 году предпринял попытку количественного анализа звуковых сочетаний в индоевропейских языках. Его целью было выявить частотные закономерности, что можно считать одной из первых форм частотного анализа языкового материала. Вслед за ним, в конце XIX века, Фридрих В. Кэдинг составил первый масштабный частотный словарь немецкого языка, в котором было проанализировано почти 11 миллионов словоформ. Этот словарь отличался от предыдущих попыток не только объёмом, но и многоуровневым анализом: Кэдинг учитывал частотность не только слов, но и слогов, морфем и буквосочетаний, что позволяло делать

выводы о словообразовательных тенденциях немецкого языка [4, 18].

Следующим важным этапом стало развитие теоретической базы частотной лингвистики в XX веке. В 1950-е годы П. Менцерат ввёл различие между объективной частотностью, или частотой фактического употребления (*Gebrauchshäufigkeit*), и системной частотностью, или частотой как регулярностью, повторяемостью внутри языка (*systematische Frequenzstatistik*) [3, 277].

Он стремился не просто фиксировать употребление, а выявить устойчивые фонетические и структурные типы слов, классифицируя словарный состав по длине, структуре и позициям ударения [5, 7].

Корпусная лингвистика как самостоятельное направление начала складываться в 1960-х годах, когда появились первые корпусные проекты, основанные на частотной организации текстов – сначала на английском и немецком языках. Однако лишь в первой половине 1990-х годов она получила признание как отдельная дисциплина. В это время было закреплено понятие корпуса как машиночитаемой базы, репрезентирующей язык в его естественном употреблении.

Современное понимание корпуса во многом сформировалось благодаря Джону Синклеру, который определял его как собрание текстов, отобранных для представления языкового разнообразия в естественной форме [Sinclair, 1991]. Вслед за ним М. Стаббс подчёркивал целевую установку корпусов: они создаются для исследования и обучения [7, 239–240].

Сегодня национальные корпуса – такие как Национальный корпус русского языка, Британский национальный корпус, Американский национальный корпус и другие – представляют собой богатые ресурсы с десятками и сотнями миллионов словоупотреблений. Эти базы включают как письменные, так и устные тексты различных жанров, стилей и региональных вариантов, что делает возможным масштабное и точное лингвистическое исследование. Корпусная лингвистика, опираясь на эти ресурсы, стала важнейшим направлением современной лингвистики, объединившим точность статистики и глубину филологического анализа.

Развитие корпусной лингвистики в Японии

Для японского языка, обладающего уникальной письменной системой и сложной грамматической структурой, создание качественных корпусов представляло особую важность и одновременно значительные трудности. На сегодняшний день японские корпусные исследования являются мощным инструментом для анализа языка. Корпусы японского языка, созданные различными исследовательскими центрами, способствуют развитию лингвистики, особенно в таких областях, как историческое языкознание, изучение разговорной речи и автоматическая обработка текстов.

Несмотря на то, что в отечественной лингвистике уже давно проводятся различные исследования в области корпусной лингвистики, корпусы японского языка до сих пор являются мало исследованными и редко употребляемыми инструментами для русских ученых. Это отчасти связано с тем, что многие японские корпусы являются платными, не все понимают, где их искать и как с ними работать. С целью донести до потенциального пользователя информацию о наличии японских корпусов и способах их использования, в 2022 году Корецкой И.Л. была опубликована статья «Корпусы Государственного института японского языка и лингвистики». В этой статье автор знакомит читателя с основными корпусами японского языка и их ключевыми чертами [8, 81–100.].

20 декабря 2023 года Национальный институт японского языка и лингвистики (国立国語　研究所, National Institute of Japanese Language and Linguistics, NINJAL) отпраздновал свое 75-летие [https://kotobaken.jp/info/news-231220-01/ (официальный сайт NINJAL)]. По мнению И.Л. Корецкой, его по праву считают организацией, заложившей основы корпусной лингвистики в Японии [8, 82].

Далее представлена эволюция японских языковых корпусов в период с 1960-х годов по настоящее время. Особое внимание уделяется технологическим аспектам создания и обработки корпусов.

Первые этапы создания корпусов японского языка (1950 – 1970-е годы)

В послевоенные годы Национальный институт японского языка (国立国語研究所) стал настоящим полигоном для лингвистических инноваций. Молодые исследователи сознательно отошли от традиционных для гуманитарных наук методов индивидуальной работы, сделав ставку на коллективные проекты. Такой подход позволил осуществлять масштабные исследования, невозможные в рамках классической филологии, и фактически являлся пионерским этапом разработки корпусной лингвистики в Японии.

Начав с ручных методов обработки данных, например, кропотливого анализа месячного выпуска газеты «Асахи» (1952) с использованием бумажных карточек (рис.1), институт быстро перешел к передовым технологиям. Уже в 1950-х здесь начали применять звукозаписывающую аппаратуру для изучения диалектов, а в 1960-х первыми в гуманитарной сфере внедрили компьютерные технологии. Особое значение имело сочетание двух подходов: статистического анализа больших массивов данных (как в первом лексическом исследовании газетного языка) и детального описания конкретных языковых явлений (например, фундаментальный труд о частицах и вспомогательных глаголах 1951 года) [10, 31].

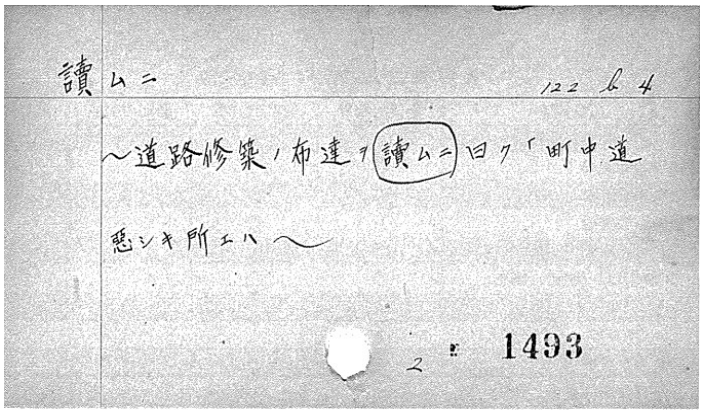

図１　手書きカードの例（『郵便報知』）

Рис.1. Пример карточки для обработки данных

Эти методологические прорывы привели к созданию уникальных исследовательских проектов. В 1956 году сотрудники института стали инициаторами основания Общества количественной лингвистики – одной из первых в мире организаций такого профиля. Последовательное внедрение компьютерных технологий (начиная с 1960-х) завершилось созданием современных корпусов, включая знаменитый Сбалансированный корпус письменного японского языка [10, 30].

Таким образом, институт не просто адаптировал новые методы, но и создал целую школу японской корпусной лингвистики, сочетающую строгую эмпирическую базу с глубоким филологическим анализом. Эта традиция продолжается в современных проектах, сохраняя актуальность подходов, разработанных несколько десятилетий назад.

Можно сказать, что основная история японских языковых корпусов начинается в 1960-х годах, когда Национальный институт японского языка (NINJAL) приступил к систематическому сбору и анализу языковых данных. NINJAL начал статистическое исследование японской лексики, став одним из пионеров в этой области. Исследователи собирали большие объемы данных из журналов и газет для анализа лексики и грамматики. Этот этап можно назвать «пионерским периодом» корпусной японской лингвистики. Например, в 1962 году был проведен проект «Анализ 90 журналов» (雑誌九十種調査), который стал первым масштабным проектом по созданию репрезентативной выборки японских текстов. Методологически он опережал свое время. Там использовалась стратифицированная выборка по жанрам и тематикам, была разработана система кодирования текстовых особенностей, впервые применены статистические методы анализа частотности [10, 8].

В отличие от западных стран, где корпусная лингвистика развивалась преимущественно в университетской среде, в Японии эту работу возглавило государственное научное учреждение. Первые проекты NINJAL носили ярко выраженный прикладной характер и были направлены на решение конкретных задач: стандартизация японской письменности, унификация правил использования иероглифов, создание учебных материалов для изучающих японский язык.

Тем не менее, некоторые ученые отмечают следующую проблему: несмотря на развитие цифровых технологий, корпусы оставались закрытыми, что ограничивало их влияние на научное сообщество.

Особенностью этого периода было отсутствие самого термина «корпус» – исследователи говорили о «языковых материалах» или «текстовых базах данных» [10, 6].

Миядзима Тацуо в своей работе «От обзоров словарного запаса к корпусам» рассматривает обзоры словарного состава, которые Национальный институт японского языка и лингвистики проводил с момента своего основания, где, в частности, обсуждается «Анализ 90 журналов» 1962 года, который существует и по сей день. Методологии, принятые в исследовании почти 50 лет назад, такие как «установление заглавных слов», «метод выборки» и «обеспечение репрезентативности», упомянутые Миядзима, также будут полезны в будущих исследованиях лексики и лингвистических исследованиях с использованием корпусов [11, 37]. Однако технические ограничения эпохи не позволили в полной мере реализовать потенциал этого проекта: все данные хранились на бумажных носителях, обработка велась вручную, отсутствовали стандартные процедуры архивирования.

Компьютеризация исследований (1966 год)

Новый этап был ознаменован компьютеризацией исследований в середине 1960-х годов. Профессор Университета Тохоку (東北大学) Хитоси Гото в своей статье «Корпусная лингвистика и изучение японского языка» пишет, что в это время некоторые лингвисты в Японии уже рассматривали и внедряли применение компьютеров в языковых исследованиях. На самом деле, история метрологии с использованием компьютеров в Японии

достаточно долгая. В 1966 году Национальный институт японского языка и лингвистики внедрил компьютеры и начал применять их в исследованиях японского языка, достигнув определенных успехов в исследовании иероглифов и лексики (терминологии и письма) [12, 49] Тем не менее, эти достижения оставались достоянием узкого круга специалистов. Гото пишет, что, помимо этого, ученые сталкивались с такими проблемами, как отсутствие стандартизированных форматов данных, ограниченный доступ к вычислительным ресурсам, неразвитость теоретической базы корпусной лингвистики.

На ранних этапах развития корпусной лингвистики японского языка исследователи сталкивались и с рядом уникальных технических трудностей. Одной из ключевых проблем было отсутствие единого стандарта кодировки, что затрудняло корректное отображение и хранение иероглифов в цифровом виде. Ещё одной существенной преградой стала невозможность автоматического сегментирования текста: в японском языке отсутствуют пробелы между словами, что требует предварительной морфологической обработки. Дополнительную сложность представляла агглютинативная природа языка и богатая система флексий, что делало автоматический анализ грамматической структуры крайне затруднительным [12, 49].

Для решения этих задач в 1960 – 1970-х годах были предложены различные инженерные подходы. Разрабатывались специализированные устройства для ввода иероглифов, создавались первые словари, пригодные для автоматического морфологического анализа, а хранение текстов осуществлялось с помощью перфокарт. Несмотря на технологическую ограниченность того времени, именно в этот период были заложены основы будущих достижений в области японской корпусной лингвистики [12, 49].

Западное влияние (1980 – 1990-е годы)

1980-е годы стали временем активного взаимодействия японских лингвистов с западными коллегами [12, 48]. Знаковым событием стал визит известного британского лингвиста Джеффри Лича (Geoffrey Leech) в 1984 году. Он представил японским исследователям концепции сбалансированных корпусов, методы корпусного анализа и опыт создания таких проектов, как Brown Corpus и LOB Corpus. Однако внедрение этих идей в Японии столкнулось с рядом трудностей: техническая инфраструктура была недостаточно развита, отсутствовали подготовленные кадры, а применимость западных методов вызывала скепсис среди японских исследователей [12, 47–48].

Несмотря на то, что Дж. Лич провёл в Японии несколько лекций, где подробно освещал возможности корпусной лингвистики, как отмечает Гото, лишь немногие японские учёные по-настоящему осознали значение этих идей. Таким образом, прямое влияние западных разработок на японскую лингвистику стало ощутимо лишь позднее [12, 47–48].

Ситуация начала меняться в 1990-х годах, когда персональные компьютеры стали более доступными и мощными. Это открыло возможности для создания первых электронных корпусов. В этот период появляются любительские и экспериментальные проекты, основанные на коммерческих электронных публикациях, литературных произведениях и новостных статьях. Также предпринимаются первые шаги к стандартизации форматов и структурированию данных [10, 6].

Несмотря на качественный скачок, новые электронные корпуса страдали от системных недостатков. Они часто формировались из доступных источников, не отражающих всего многообразия японского языка. Проблемы авторского права ограничивали распространение корпусов,

форматы хранения данных были несовместимыми между различными исследовательскими центрами, а слабая документация существенно затрудняла воспроизводимость исследований [12, 51].

Тем не менее, именно в этот переходный период начинается формирование ключевых подходов к созданию электронных корпусов. Наиболее значимыми проектами стали корпус газеты «Асахи», содержащий около 50 миллионов слов с ручной разметкой, корпус художественной литературы XX века и корпус научных текстов, включавший специализированную лексику [11, 37]. Техническими достижениями этого этапа можно считать разработку первых японских конкордансеров, создание алгоритмов морфологической разметки и эксперименты со статистическими методами анализа.

Характерным признаком данного периода стало закрепление термина «электронный корпус», который позже стал обозначаться просто как «корпус» [12, 52]. Однако отсутствие сбалансированных и методологически выверенных корпусов продолжало оставаться серьёзным вызовом.

Несмотря на прогресс, корпусная лингвистика 1990-х страдала от системных недостатков, таких как отсутствие баланса (корпусы отражали только доступные газетные и журнальные тексты); проблемы авторского права (ограничения на распространение); несовместимость форматов (каждый центр использовал свои стандарты); недостаток метаданных (слабая документация корпусов). Эти проблемы стали стимулом для качественного скачка в следующем десятилетии.

В начале 1990-х годов повышение производительности персональных компьютеров позволило исследователям использовать цифровые версии газетных статей и литературных произведений.

Современный этап: эпоха профессиональных корпусов (2000-е – настоящее время)

С начала 2000-х годов японская корпусная лингвистика вступила в новый этап своего развития, характеризующийся созданием репрезентативных, масштабных и методологически обоснованных корпусов. Ключевым событием стало начало проекта Balanced Corpus of Contemporary Written Japanese (BCCWJ), запущенного Национальным институтом японского языка и лингвистики (NINJAL) в 2006 году и завершённого в 2011 году. Этот проект стал не только техническим достижением, но и научной вехой, заложив основы «третьей эпохи» японской корпусной лингвистики. BCCWJ (Balanced Corpus of Contemporary Written Japanese) представляет собой корпус, включающий около 100 миллионов слов, отобранных по принципу репрезентативности из различных источников современного письменного японского языка – книг, газет, журналов, интернет-ресурсов и других. Методология построения корпуса основывалась на чётких критериях отбора текстов, жанровой стратификации и учёте социолингвистических факторов. Важной особенностью стало использование XML-формата с расширенной аннотацией, что обеспечивало высокую степень формализованностии гибкость при последующем анализе [13, 2].

Корпус стал неоценимым ресурсом для изучения синтаксических структур, коллокаций, грамматических паттернов и частотных характеристик. Он активно используется как в фундаментальных лингвистических исследованиях, так и в прикладных областях – от лексикографии до преподавания японского языка.

Параллельно с BCCWJ получили развитие специализированные корпусные проекты. Одним из наиболее значимых

является Corpus of Spontaneous Japanese (CSJ) – масштабная коллекция разговорной речи, включающая около 700 часов аудиозаписей, транскрибированных с фонетической аннотацией. CSJ стал важнейшим источником для анализа устной коммуникации, интонационных структур и прагматических маркеров [13, 5-6].

Также активно развиваются исторические корпуса, охватывающие тексты периодов Эдо, Мэйдзи и Тайсё. Эти проекты требуют специфических методов цифровизации, включая распознавание рукописей и обработку устаревших графических форм. Кроме того, создаются корпусы для изучающих японский как иностранный язык (например, I-JAS), которые позволяют анализировать ошибки, типичные для учащихся, и разрабатывать адаптивные методики преподавания [8, 90].

Современные технологии существенно расширили возможности корпусной лингвистики. Использование методов глубокого обучения и алгоритмов автоматической аннотации повысило точность морфологического и синтаксического анализа. Появление удобных веб-интерфейсов и открытых API обеспечило доступ к корпусам широкому кругу исследователей и разработчиков лингвистических приложений. Японские корпуса стали важным компонентом экосистемы NLP, играя ключевую роль в развитии машинного перевода, анализа текста, голосовых помощников и других интеллектуальных систем.

Обратимся к рассмотрению наиболее востребованных корпусов японского языка (см. Рис. 2). https://www.bunka.go.jp/seisaku/bunkashingikai/kokugo/gengo/gengo_01/pdf/94081101_05.pdf

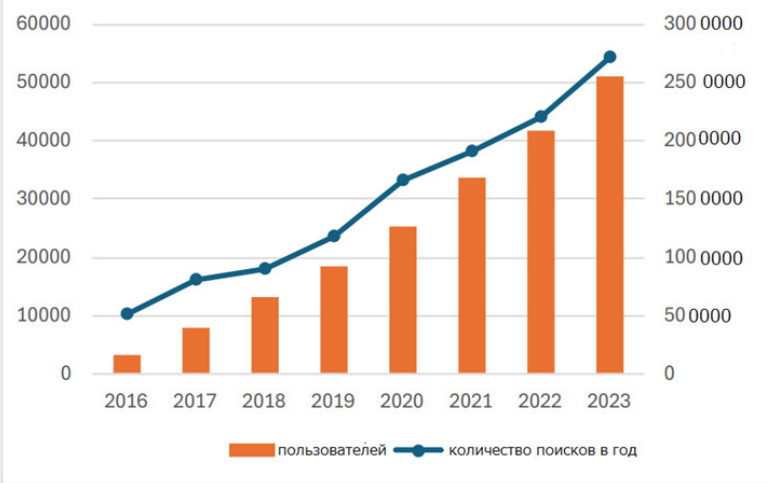

Рис.2. Количество пользователей бесплатной онлайн-системы корпуса «Чūпагон» и количество запросов в год

Некоторые проблемы при разработке корпуса BCCWJ

Создание японского корпуса, особенно такого масштабного проекта, как «Современный японский корпус письменной речи» (BCCWJ), сопровождалось рядом сложностей, отражающих как языковые, так и организационно-технические аспекты.

Одной из главных лингвистических проблем стало наличие вариативности в написании слов (表記のゆれ). В японском языке омонимы с разными значениями записываются разными иероглифами. Например, toru「取る」(брать),「撮る」(фотографировать),「採る」(собирать), разными окончаниями (送り仮名), либо с использованием хираганы и катаканы.

Особенно сильно вариативность проявляется в глаголах, где также возможны омонимичные формы (異字同訓). Это делает автоматическую обработку и категоризацию данных более трудоёмкой и требует сложных правил нормализации.

Кроме того, имелись различия между типами источников, включённых в корпус. Например, официальные документы и газеты следуют строгим стандартам написания, что ограничивает вариативность, тогда как блоги, веб-форумы и книги позволяют более свободное выражение, что ведёт к росту нестандартных написаний. Это затрудняет создание сбалансированного и репрезентативного корпуса.

Также возникали технические и правовые трудности. Требовалось осуществить сложную работу по выборке текстов, в том числе с применением стратифицированной случайной выборки, обеспечить согласование авторских прав, а также провести разметку и морфологический анализ с высокой степенью точности. Для этого разрабатывались специализированные инструменты и электронные словари, такие как UniDic [14, 2].

Наконец, сам процесс обработки данных (оцифровка, XML-разметка, корректура и аннотирование) потребовал значительных усилий, координации между исследовательскими группами и длительного времени, что также стало вызовом для организаторов проекта [14, 2].

Сбалансированный корпус современного японского языка (BCCWJ – Balanced Corpus of Contemporary Written Japanese)

BCCWJ является одним из крупнейших корпусов письменного японского языка, включающим книги, журналы, газеты и интернет-ресурсы. Этот корпус широко используется в исследованиях современной лингвистики, поскольку охватывает различные стили и регистры японского письма (https://clrd.ninjal.ac.jp/bccwj/index.html).

Корпус спонтанной японской речи (CSJ – Corpus of Spontaneous Japanese)

Этот корпус предназначен для изучения устной речи и содержит транскрибированные аудиозаписи различных разговорных ситуаций. Он используется в исследованиях разговорной речи, фонетики, прагматики и автоматического распознавания речи.

Интернет-корпус NINJAL (NWJC – NINJAL Web Japanese Corpus)

Этот корпус содержит огромный объем текстов, собранных из интернета, что позволяет изучать новейшие языковые тенденции, такие как появление неологизмов и изменения в структуре предложений.

Международный учебный корпус японского языка как иностранного (I-JAS – International Corpus of Japanese as a Second Language)

Этот корпус представляет собой собрание текстов, написанных изучающими японский язык иностранцами. Он используется в преподавании и исследовании процессов овладения японским языком.

Методика корпусных исследований в современной японской лингвистике

Корпусная лингвистика радикально изменила облик современной японской лингвистики, особенно с конца 1990-х годов, когда начали появляться масштабные электронные корпусы, такие как Balanced Corpus of Contemporary Written Japanese (BCCWJ) и Corpus of Spontaneous Japanese (CSJ). Эти ресурсы не просто добавили новые инструменты в арсенал исследователей, но они потребовали переосмысления самих методов, целей и подходов к языковому анализу [10, 5].

До появления корпусов японская лингвистика в значительной степени опиралась на интуитивный анализ языковых примеров, зачастую искусственно сконструированных. С корпусами исследователи получили возможность опираться на реальные, эмпирически зафиксированные данные. Это означало переход от вопроса «как это должно быть?» к вопросу «как это на самом деле используется?» Например, ранее предполагалось, что конструкция 「～ことができる」 («мочь сделать что-то») является более формальной и потому более употребительной в официальной письменной речи. Однако анализ BCCWJ (Balanced Corpus of Contemporary Written Japanese) показал, что в газетных текстах эта форма встречается реже, чем альтернативная конструкция с потенциалом глагола (напр. 書ける), особенно в статьях, ориентированных на массового читателя. Это позволило уточнить стилистическую маркированность обеих конструкций на основе частотных данных, а не только интуиции.

Ещё одним ярким примером является изучение выражений вежливости и прагматических маркеров. Ранее считалось, что такие формы, как 「～でございます」 или「お～いたします」, являются устоявшимися клише формального японского языка. Однако корпус CSJ, собравший обширные данные спонтанной устной речи, показал, что в ситуациях формального общения (например, презентации на конференциях) говорящие часто прибегают не к формальным выражениям, а к более гибким и даже разговорным формам, например, 「～っていうのは」 вместо 「～と申しますのは」. Это позволило сделать вывод о более тонком и гибком понимании японской вежливости, как ситуативной, а не фиксированной категории.

Кроме того, корпусные данные изменили представления о частотности лексики и её преподавании. Например, многие учебники ранее уделяли значительное внимание таким словам, как 例えば («например»), 因って («вследствие») или 然しながら («однако»), предполагая их высокую частотность в формальной письменной речи. Но анализ BCCWJ (Balanced Corpus of Contemporary Written Japanese) показал, что некоторые из них (например, 然しながら) на самом деле используются крайне редко, и их частотность уступает более простым выражениям вроде でも или だから, даже в текстах академического или официального стиля. Это привело к пересмотру учебных материалов, словарей и систем уровней владения языком.

Таким образом, корпусная лингвистика в Японии способствовала переходу от нормативной и теоретической модели языка к описательной, эмпирически обоснованной. Она изменила не только методы анализа, но и само понимание языка как динамической, контекстуально обусловленной системы.

О заимствованиях из китайского и английского языков

Поскольку до сих пор исчерпывающей статистики в области письменного японского языка не представлено, приходится использовать данные выборочных лексических исследований. Например, для ана-

лиза распределения типов слов (goshu, 語種) до сих пор часто ссылаются на результаты «Исследования 70 журналов». Ниже, вместе с данными нового «Исследования 70 журналов», приведены относительные частоты слов (рис. 3). В «Исследовании 70 журналов» анализировался только основной текст (без рекламы). В «Исследовании 70 журналов» представлены данные как для основного текста, так и для рекламы, но здесь используется только основной текст для сопоставимости. В последние годы kango (слова китайского происхождения) превзошли по частоте wago (исконно японская лексика). Заметно возросло количество gairaigo (заимствований из других языков, кроме китайского) [11, 40].

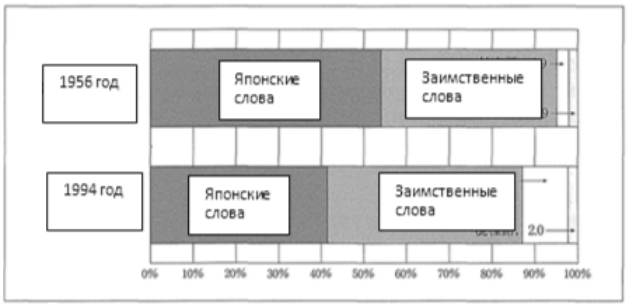

Рис. 3. Соотношение типов слов от общего количества слов в журналах

позволяют анализировать эволюцию языка на основе реальных текстов, а не только теоретических моделей. Corpus of Everyday Japanese Conversation и Corpus of Spontaneous Japanese стали незаменимыми инструментами для изучения особенностей устного общения, от интонации и пауз до жаргона и стилистических вариаций.

Наконец, современные корпуса лежат в основе автоматической обработки японского языка (NLP), обеспечивая развитие машинного перевода, голосовых ассистентов и других технологий. По данным Министерства Японской Культуры, корпусная лингвистика, лежащая в основе современных технологий обработки естественного языка, играет ключевую роль в развитии систем искусственного интеллекта, что определяет лицо будущего Японии (https://www.bunka.go.jp/seisaku/bunkashingikai/kokugo/gengo/gengo_02/pdf/94091101_02.pdf) (официальный сайт Министерства Японской Культуры).

Заключение

Развитие корпусной лингвистики в Японии значительно изменило подходы к изучению японского языка. Создание и анализ различных корпусов позволили исследователям глубже понять его структуру, историю и динамику развития. Эти данные используются не только в академической среде, но и в практических приложениях, таких как автоматический перевод, анализ текста с помощью искусственного интеллекта, преподавание японского языка.

Корпусные исследования кардинально изменили методы изучения японского языка. В области исторической лингвистики такие корпуса, как Corpus of Everyday Japanese Conversation,

Таким образом, корпусная лингвистика не только расширила научные горизонты, но и создала практическую базу для цифровых инноваций. Использование больших массивов текстов позволяет моделям машинного обучения выявлять статистические закономерности языка, строить синтаксические и семантические структуры, и, в конечном счете, приближаться к нормам естественной языковой компетенции.

Библиографический список

1. *Захаров В.П.* Корпусная лингвистика: учебник / В. П. Захаров, С. Ю. Богданова. – Иркутск: ИГЛУ, 2011. – 161 с.
2. *Сысоев П.В.* Лингвистический корпус в методике обучения иностранным языкам // Язык и культура. – 2010. –№ 1(9). – С. 99– 111.
3. *Чугаева Т.Н., Байбурова О.В., Вахотин А.А., Дмитриева (Мякотникова) С.Ю.* Сопоставление результатов лингвостатистического анализа перцептивных типов русского и английского слова (на материале НКРЯ, БНК, АНК) // Теоретическая и прикладная лингвистика. 2019. Т. 5, № 3. С. 273–291.
4. *Kaeding F.W.*: Häufigkeitswörterbuch der deutschen Sprache.Selbstverlag [Text] / Kaeding F.W // Linguistics and Linguistic Theory. –2009. – 5 (1). – S. 1–26.
5. *Menzerath, P.* Architektonik des deutschen Wortschatzes. – Bonn, 1954.
6. *Sinclair J.* Corpus Concordance Collocation [Text] / J. Sinclair. – OUP, 1991. – 197 p.
7. *Stubbs M.* Texts, corpora and problems of interpretation: A response to Widdowson [Text] / M. Stubbs // Applied Linguistics. – 2001. – Vol. 22. – Issue 2. – P. 149–172.
8. *Корецкая И.Л.* Корпусы Государственного института японского языка и лингвистики // Rhema. Рема. – 2022. – № 4. – С. 81–100. – DOI: 10.31862/2500-2953-2022-4-81-100.
9. Официальный сайт Национального Института Японского Языка (National Institute for Japanese Language and Linguistics) https://kotobaken.jp/info/ news-231220-01/
10. *Маруяма Такэхико, Таномура Тадахару* (2019). Корпусная лингвистика японского языка: перспективы исследования [コーパス日本語学の射程]. Репозиторий Национального института японского языка (国立国語研究所学術情報リポジトリ). DOI: 10.15084/00002179.
11. *Миядзима Тацуо.* От лексических исследований к корпусу [語彙調査からコーパスへ] // Репозиторий Национального института японского языка (国立国語研究所学術情報リポジトリ). – 2019. – 25 марта. – DOI: 10.15084/00002181.
12. *Гото Хитоси*. Лингвистика корпусов и исследования японского языка // Наука о японском языке (Nihongo Kagaku). – 2007. – №22. – С. 47–58. – URL: https://core.ac.uk/download/pdf/234727466.pdf).
13. *Маэкава Кикуо*. Проект «Японский корпус» в рамках приоритетных исследований: цели, текущие результаты и перспективы [特定領域研究「日本語コーパス」－目標,進捗状況,そして夢－] // Национальный институт японского языка. – Отдел исследований и разработок (Dept. Lang. Res., National Institute for Japanese Language).
14. *Maekawa, K.* (Year). Priority-area «Japanese Corpus» project: Goals, progress, and dreams. Department of Language Research, National Institute for Japanese Language. URL: https://www2.ninjal.ac.jp/kikuo/tokutei07ws_km1.pdf
15. Американский национальный корпус. [Электронный ресурс]. URL: http://www.americannationalcorpus.org (дата обращения: 26.03.2011).
16. Британский национальный корпус. [Электронный ресурс]. URL: http://www.natcorp.ox.ac.uk (дата обращения: 14.07.2011).
17. Венгерский национальный корпус. [Электронный ресурс]. URL: http://corpus.nytud.hu/mnsz/index_eng.html (дата обращения: 30.07.2015).
18. Корпус немецкого языка. [Электронный ресурс]. URL: http://www.dwds.de (дата обращения: 30.07.2015).
19. Корпус современного китайского языка. [Электронный ресурс]. URL: http://www.cncorpus.org (дата обращения: 30.07.2015).
20. Корпус французского языка. [Электронный ресурс]. URL: http://sites.univprovence.fr/delic/corpus/index.htm (дата обращения: 30.07.2015).

HISTORY AND CURRENT STATE OF CORPUS LINGUISTICS IN JAPAN

Zhdanova Y.V. [1,2]

[1] *Perm Federal Research Center of the UB RAS*
[2] *Perm State National Research University*

For citation:
Zhdanova Y.V. History and current state of corpus linguistics in Japan // Perm Federal Research Center Journal Perm Federal Research Center Journal. – 2025. – № 2. – P. 19–31. https://doi.org/10.7242/2658-705X/2025.2.2

The rapid development of corpus linguistics in the 1990s fundamentally changed the context of linguistic research, providing European scholars with new opportunities to work with vast amounts of language data. In Japanese linguistics, the fascinating process of creating a national language corpus has faced significant challenges for various reasons. However, in the Russian language linguistic literature, sufficient information on the state of Japanese language corpora is currently conspicuously scarce. This lacuna was partially addressed in an article by I.L. Koretskaya (2022), which attempts to outline the main Japanese language corpora available today. Nevertheless, the history of the development of Japanese corpora and its current state require more thorough description and analysis because of the urgency of solving intralinguistic issues in the Japanese language. This article traces key aspects of the history and development of Japanese language corpora, and highlights some of the difficulties that Japanese linguists have to overcome at different stages of creating a national corpus of the Japanese language.

Keywords: Japanese language corpus, development of corpus linguistics, Japanese corpus linguistics, corpus of tests, corpus of written Japanese, corpus of spontaneous speech, International teaching Corpus of Japanese as a Foreign Language.

Сведения об авторах

Жданова Юлия Владимировна, кандидат филологических наук, старший преподаватель кафедры иностранных языков и философии, Пермский федеральный исследовательский центр УрО РАН (ПФИЦ УрО РАН), 614000, г. Пермь, ул. Ленина, д.13А, ассистент кафедры теоретического и прикладного языкознания, Пермский государственный национальный исследовский университет (ПГНИУ), 614013, г. Пермь, ул. Букирева, 15; e-mail: yukayokyok@gmail.com

Материал поступил в редакцию 18.04.2025