

МРНТИ 16.21.07

<https://doi.org/10.26577/EJPh202220262>А.Ф. Аляхимет^{1*}, Ж.М. Уматова²¹Казахский национальный университет имени аль-Фараби, Алматы, Казахстан²Синьцзянский университет, Урумчи, КНР

*e-mail: alyakhimet.aidana@gmail.com

ТИПОЛОГИЧЕСКАЯ ОБУСЛОВЛЕННОСТЬ АРХИТЕКТУРЫ НАЦИОНАЛЬНЫХ ЯЗЫКОВЫХ КОРПУСОВ: СОПОСТАВИТЕЛЬНЫЙ АНАЛИЗ

В статье рассматриваются национальные языковые корпуса как стратегические ресурсы современной корпусной лингвистики и цифровых гуманитарных наук. Актуальность исследования обусловлена необходимостью типологически ориентированного анализа корпусных ресурсов, учитывающего структурные особенности языка при их проектировании. Целью работы является сопоставительный анализ Национального корпуса казахского языка, Национального корпуса русского языка и Турецкого национального корпуса с точки зрения влияния типологической структуры языка на архитектуру корпуса и уровни лингвистической аннотации. Методологическую основу исследования составляют описательный и сравнительный методы, элементы качественного корпусного анализа, а также параметрическое сопоставление объемов, жанрового состава и уровней морфологической, морфосинтаксической и семантической разметки. В результате выявлено, что для корпусов агглютинативных языков приоритетное значение имеет детализированная морфологическая аннотация, обеспечивающая корректную сегментацию аффиксов и лемматизацию, тогда как для флективного русского языка ключевым является многоуровневое морфосинтаксическое и семантическое описание. Установлена зависимость архитектуры национального корпуса от типологических характеристик языка. Научная ценность работы заключается в формировании типологически ориентированной модели сопоставления национальных корпусов. Практическое значение результатов состоит в возможности их использования при дальнейшем развитии Национального корпуса казахского языка, а также в образовательных и прикладных задачах автоматической обработки естественного языка.

Ключевые слова: корпусная лингвистика, национальный языковой корпус, типология языков, морфологическая аннотация, корпусное моделирование.

A. Alyakhimet^{1*}, Zh. Umatova²¹Al-Farabi Kazakh National University, Almaty, Kazakhstan²Xinjiang University, Ürümqi, China

*e-mail: alyakhimet.aidana@gmail.com

Typological Determination of the Architecture of National Language Corpora: a Comparative Analysis

The article examines national language corpora as strategic resources of modern corpus linguistics and digital humanities. The relevance of the study is determined by the need for a typologically oriented comparative analysis of national corpora that takes into account the structural characteristics of languages in corpus design. The aim of the research is to conduct a comparative analysis of the National Corpus of the Kazakh Language, the National Corpus of the Russian Language, and the Turkish National Corpus in order to determine how typological language structure influences corpus architecture and levels of linguistic annotation. The methodological framework includes descriptive and comparative methods, elements of qualitative corpus analysis, and parametric comparison of corpus size, genre composition, and levels of morphological, morphosyntactic, and semantic annotation. The results demonstrate that corpora of agglutinative languages prioritize detailed morphological annotation ensuring accurate morpheme segmentation and lemmatization, whereas for the inflectional Russian language multi-level morphosyntactic and semantic annotation plays a central role. The study substantiates the dependence of corpus architecture on typological language characteristics. The scientific contribution of the research lies in proposing a typologically grounded analytical model for comparing national language corpora. The practical significance of the findings consists in their applicability to the further development of the National Corpus of the Kazakh Language, as well as to educational practice and applied tasks of natural language processing.

Keywords: corpus linguistics, national language corpus, language typology, morphological annotation, corpus architecture.

А.Ф. Аляхимет^{1*}, Ж.М. Уматова²

¹Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан

²Шыңжаң университеті, Үрімші, ҚХР

*e-mail: alyakhimet.aidana@gmail.com

Ұлттық тілдік корпустар архитектурасының типологиялық детерминациясы: салыстырмалы талдау

Мақалада ұлттық тілдік корпустар заманауи корпус лингвистикасы мен цифрлық гуманитарлық ғылымдардың стратегиялық ресурстары ретінде қарастырылады. Зерттеудің өзектілігі тілдің құрылымдық ерекшеліктерін ескере отырып корпус ресурстарын жобалауда типологиялық бағытталған талдау жүргізу қажеттілігімен айқындалады. Жұмыстың мақсаты – тілдің типологиялық құрылымының корпус архитектурасы мен лингвистикалық аннотация деңгейлеріне ықпалын анықтау тұрғысынан Қазақ ұлттық корпусы, Орыс тілінің ұлттық корпусы және Түрік ұлттық корпусының салыстырмалы талдауын жүргізу. Зерттеудің әдіснамалық негізін сипаттамалық және салыстырмалы әдістер, сапалық корпус талдауының элементтері, сондай-ақ көлемі, жанрлық құрамы және морфологиялық, морфосинтаксистік, семантикалық таңбалау деңгейлерін параметрлік салыстыру құрайды. Зерттеу нәтижесінде агглютинативті тілдер корпустары үшін аффикстердің дұрыс сегментациясын және лемматизацияны қамтамасыз ететін егжей-тегжейлі морфологиялық аннотацияның басым мәнге ие екендігі, ал флексивті орыс тілі үшін көпдеңгейлі морфосинтаксистік және семантикалық сипаттаудың шешуші рөл атқаратыны анықталды. Ұлттық корпус архитектурасының тілдің типологиялық сипаттамаларына тәуелділігі айқындалды. Зерттеудің ғылыми құндылығы ұлттық корпустарды салыстырудың типологиялық бағытталған моделін қалыптастырумен сипатталады. Алынған нәтижелердің практикалық маңызы оларды Қазақ ұлттық корпусын одан әрі дамытуда, сондай-ақ білім беру және табиғи тілді автоматты өңдеу салаларындағы қолданбалы міндеттерді шешуде пайдалануға болатындығында көрінеді.

Түйін сөздер: корпус лингвистикасы, ұлттық тілдік корпус, тіл типологиясы, морфологиялық аннотация, корпус архитектурасы.

Введение

В современной филологической парадигме корпусная лингвистика функционирует как эмпирически ориентированная исследовательская модель, основанная на анализе структурированных цифровых массивов языковых данных (Sinclair, 1991, p. 4; McEnery Hardie, 2012, p. 2). В отличие от традиционных описательных подходов, корпусная методология предполагает работу с репрезентативными текстовыми коллекциями, организованными в виде многоуровневых цифровых систем. Национальные языковые корпуса в данном контексте представляют собой институционализированные цифровые инфраструктуры, обеспечивающие формализованное описание языка в синхронии и диахронии (Biber, Conrad, Reppen, 1998, p.15; Ide, Pustejovsky, 2017, p. 8). Их функциональное значение распространяется на лексикографию, образовательные технологии, задачи автоматической обработки естественного языка и реализацию языковой политики в цифровой среде (Фазылжанова, 2021, б. 16).

Рассмотрение национальных корпусов сквозь призму языковой типологии позволяет выявить зависимость архитектурных решений

от структурных параметров конкретного языка. Проектирование корпуса не является нейтральной технической процедурой: глубина аннотации, выбор уровней разметки и конфигурация подкорпусов детерминируются морфологической организацией языковой системы (Comrie, 1989, p. 31; Haspelmath, Sims, 2010, p. 9). Несмотря на значительный объем исследований в области корпусного моделирования, сопоставительные работы, рассматривающие национальные корпуса как типологически обусловленные системы, представлены ограниченно (McEnery, Xiao, Tono, 2006, p. 18).

Казахский и турецкий языки, относящиеся к агглютинативному типу, характеризуются линейной аффиксацией и прозрачной морфемной структурой слова, что требует высокой точности сегментации и алгоритмов лемматизации при корпусной (Haspelmath, Sims, 2010: 47; Мадиева, Уматова, 2020, p. 49). Русский язык, представляющий флексивную модель, демонстрирует синкретизм грамматических значений и вариативность синтаксических связей, что предопределяет необходимость многоуровневого морфосинтаксического и семантического описания (Богданов, Плунгян, Сичинава, 2005, с. 8; Ляшевская, 2022, с. 42).

Цель настоящего исследования состоит в выявлении степени влияния типологической структуры языка на архитектуру национального корпуса и приоритетные уровни его аннотации на материале казахского, русского и турецкого языков.

Научная новизна работы определяется разработкой типологически ориентированной модели сопоставления национальных корпусов, в рамках которой корпус рассматривается как структурно детерминированная система цифрового представления языка.

Материалы и методы

Эмпирическую основу исследования составили три национальных корпусных ресурса: Национальный корпус казахского языка, Национальный корпус русского языка (в редакции НКРЯ 2.0, 2024 г.) и Turkish National Corpus (TNC). Отбор указанных корпусов обусловлен их институциональным статусом официальных репрезентативных цифровых ресурсов, предназначенных для системного описания языковых структур, а также типологической неоднородностью представляемых языков. Казахский и турецкий языки относятся к агглютинативному морфологическому типу, русский язык – к флективному, что обеспечивает возможность выявления корреляции между морфологической организацией языка и конфигурацией корпусной архитектуры.

В качестве объекта анализа рассматриваются корпусные архитектуры как цифровые модели языковой системы. Предметом исследования выступают структурные и аннотационные параметры корпусов, включая уровни лингвистической разметки и принципы организации текстовой базы. Материал анализировался на основе официальных интерфейсов корпусов, опубликованных описаний их архитектуры и аннотационных стандартов, а также документации, отражающей функциональные возможности поисковых механизмов.

В исследовании учитывались следующие параметры: объём текстового массива; жанрово-стилистическая стратификация; наличие диахронических подкорпусов; глубина морфологической сегментации; степень формализации морфосинтаксической аннотации; характер лемматизации; интеграция корпуса с внешними лингвистическими ресурсами и стандартами (в том числе зависимостными моделями размет-

ки). Совокупность указанных критериев позволяет рассматривать корпус не как совокупность текстов, а как структурированную цифровую систему репрезентации языковых единиц и отношений.

Постановка исследования обусловлена недостаточной разработанностью типологически ориентированного сопоставления национальных корпусов. Несмотря на наличие обширных описательных исследований отдельных проектов, параметрический анализ корпусных архитектур в зависимости от морфологического типа языка представлен фрагментарно. Это определило необходимость формализации методологического инструментария.

В рамках исследования сформулированы следующие исследовательские вопросы:

- коррелирует ли морфологический тип языка с конфигурацией корпусной архитектуры;
- каким образом морфологическая организация языковой системы влияет на иерархию аннотационных уровней;
- является ли модель национального корпуса универсальной или типологически детерминированной.

Рабочая гипотеза основывается на положении о типологической обусловленности корпусной архитектуры, согласно которому морфологическая структура языка предопределяет глубину морфемной сегментации, конфигурацию аннотационных слоёв и стратегию цифрового моделирования языковых данных.

Методологическую основу исследования составляют принципы корпусной лингвистики, сравнительно-типологического анализа и структурного моделирования языковых систем. В работе применялись описательный метод (для систематизации архитектурных характеристик корпусов), сравнительно-типологический метод (для выявления структурных различий и сходств), а также параметрический анализ, предусматривающий сопоставление корпусов по формализованному набору критериев. Элементы качественного корпусного анализа использовались для интерпретации особенностей аннотационных решений в контексте типологических характеристик языка.

Современные корпусные исследования характеризуются интеграцией количественно ориентированных методов анализа языковых данных. В англоязычной традиции корпус рассматривается как формализованная эмпирическая база, допускающая применение статисти-

ческих инструментов и моделей распределения языковых единиц (Brezina, 2018, p. 25; Gries, 2019, p. 18). Использование параметрического сопоставления в настоящем исследовании соотносится с указанным направлением и позволяет рассматривать корпусную архитектуру как структурно моделируемую систему.

Предложенный методический подход обеспечивает переход от дескриптивного описания отдельных проектов к объяснительной модели, рассматривающей национальный корпус как типологически детерминированную цифровую репрезентацию языковой системы.

Обзор литературы

Развитие корпусной лингвистики связано с переходом от интроспективного описания языка к эмпирически верифицируемым моделям анализа, основанным на систематизированных текстовых массивах. В классических исследованиях J. Sinclair (1991) корпус определяется как структурированная репрезентация языка, предназначенная для выявления устойчивых закономерностей употребления. Подчеркивается, что корпус должен быть функционально сбалансированным и отражать реальные коммуникативные практики, а не представлять собой произвольный набор текстов. Тем самым формируется понимание корпуса как исследовательского инструмента, ориентированного на объективируемые языковые данные.

В работах D. Biber, S. Conrad и R. Reppen (1998) корпусная методология получает количественное обоснование. Авторы демонстрируют, что репрезентативность определяется не только объемом, но и продуманной жанровой структурой и функциональным распределением текстов. Развитие этих положений прослеживается в исследованиях Т. McEnery и А. Hardie (2012), где корпусная лингвистика представлена как комплексная методологическая система, включающая теоретические и прикладные аспекты анализа. Исследователи подчеркивают, что глубина аннотации определяет аналитический потенциал корпуса и его применимость в задачах автоматической обработки естественного языка.

Современный этап развития корпусной лингвистики характеризуется усилением количественно ориентированных методов анализа языковых данных и интеграцией статистических моделей в исследовательскую практику

(Gries, 2021). Корпус в англоязычной традиции рассматривается как многоуровневая цифровая модель языковой системы, включающая взаимосвязанные аннотационные слои различной степени формализации, что отражено в работах Т. McEnery и V. Brezina (McEnery, Brezina, 2022, p. 14). Эти подходы усиливают интерпретационный потенциал корпусных исследований и расширяют возможности типологического сопоставления языков.

Проблематика аннотационных моделей получила развитие в трудах N. Ide и J. Pustejovsky (2017), где разметка рассматривается как формализованная система кодирования лингвистических знаний. Международные инициативы, в частности Universal Dependencies (Nivre et al. 2020), демонстрируют стремление к унификации морфосинтаксической аннотации в многоязычном пространстве. Однако универсальные стандарты неизбежно требуют адаптации к структурным особенностям конкретных языков.

В этом контексте особое значение приобретает языковая типология. В. Comrie (1989) и М. Haspelmath, А. Sims (2010) систематизируют различия между агглютинативными и флективными языками, указывая на различия в способах выражения грамматических значений. Агглютинативные языки характеризуются последовательной аффиксацией и формальной прозрачностью морфемных границ, тогда как флективные языки демонстрируют синкретизм и формальное слияние грамматических показателей. Эти различия оказывают непосредственное влияние на принципы проектирования корпусной архитектуры, глубину морфологической сегментации и необходимость синтаксической аннотации.

В рамках российской корпусной лингвистики значительный вклад в развитие корпусной теории представлен в работах С.И. Богданова, В.А. Плунгяна и Д.В. Сичинавы (2005) представлена концепция многоуровневой аннотации, объединяющей морфологическое и синтаксическое описание. Современный этап развития корпуса отражён в публикации С. О. Савчук и соавторов (2024), посвящённой модернизации платформы НКРЯ 2.0. Авторы отмечают внедрение новых алгоритмов обработки данных, расширение подкорпусов и совершенствование интерфейса аналитического доступа. Это позволяет рассматривать НКРЯ как динамично развивающуюся цифровую инфраструктуру, адаптирующуюся к современным исследовательским задачам.

Кроме того, корпус активно используется в прикладных исследованиях. Так, в работах 2020-2025 гг. анализируются современные дискурсивные практики и концепты цифровой эпохи на материале НКРЯ, что демонстрирует его востребованность в междисциплинарных исследованиях.

В казахстанской научной школе вопросы корпусного моделирования рассматриваются в контексте цифровизации национального языка и укрепления научной инфраструктуры. В исследованиях Г.Б. Мадиевой и Ж.М. Уматовой (2020) анализируются принципы формирования Национального корпуса казахского языка и подчёркивается необходимость детализированной морфологической аннотации. А.М. Фазылжанова (2021) связывает развитие корпусных технологий с задачами модернизации языковой политики.

Теоретическую основу морфологического анализа казахского языка формируют труды Қ. Жұбанова (Жұбанов, 1999, б. 15), Ә. Қайдар (Қайдар, 1998, б. 22), К. Аханова (Аханов, 2002, б. 47), в которых системно описаны закономерности словообразования и словоизменения. В исследованиях Н. Уәли (Уәли, 2012, б. 30) и Ш. Құрманбайұлы (Құрманбайұлы, 2005, б. 18) особое внимание уделяется вопросам нормализации, терминологической систематизации и стандартизации языка в цифровую эпоху. Эти положения имеют прямое отношение к корпусной практике, поскольку корпус выступает инструментом фиксации и кодификации языковой нормы.

Исследования Института языкознания им. А. Байтурсынова подчёркивают стратегическое значение Национального корпуса казахского языка как цифрового ресурса, обеспечивающего сохранение языкового наследия и развитие прикладных лингвистических технологий.

Турецкий национальный корпус (Aksan, 2012) представляет собой пример системного проектирования корпуса агглютинативного языка. В последние годы наблюдается активное развитие турецкой корпусной экосистемы, включая интеграцию TNC с web-корпусами, парламентскими корпусами и зависимостными treebank-проектами (2020-2023 гг.). Это свидетельствует о переходе от изолированной корпусной модели к комплексной цифровой инфраструктуре.

Несмотря на значительный вклад перечисленных исследований, анализ литературы выявляет определённый пробел. Большинство работ

либо посвящено теоретическим аспектам корпусной лингвистики, либо описанию отдельных национальных проектов. Системное типологическое сопоставление национальных корпусов агглютинативных и флективных языков в рамках единой аналитической модели представлено ограниченно. В частности, отсутствует параметрический анализ, позволяющий выявить зависимость архитектуры корпуса от морфологической структуры языка с учётом национальных научных традиций.

Таким образом, существующие исследования формируют теоретическую и методологическую основу, однако не предлагают целостной модели типологически ориентированного сопоставления национальных корпусов. Настоящее исследование направлено на восполнение данного пробела посредством разработки аналитической схемы, в рамках которой национальный корпус рассматривается как типологически детерминированная цифровая система, архитектура которой обусловлена структурными параметрами языка и исследовательскими приоритетами научной школы.

Результаты и обсуждение

Типологическая обусловленность архитектуры корпуса. Проведённый сопоставительный анализ национальных корпусов казахского, русского и турецкого языков подтвердил выдвинутую гипотезу о типологической детерминированности архитектуры корпусного моделирования. Установлено, что архитектурные решения, глубина и структура аннотации, а также функциональные параметры корпусов находятся в системной зависимости от морфологической организации языка.

В корпусах агглютинативных языков – казахского и турецкого – центральным структурообразующим элементом выступает морфологическая аннотация. Детализированная сегментация аффиксов, последовательное выделение словообразовательных и словоизменительных показателей и прозрачная лемматизация словоформ составляют основу аналитической модели данных корпусов. Подобный подход согласуется с положениями языковой типологии, согласно которым агглютинативные системы характеризуются формальной дискретностью морфем и линейной структурой слова (Haspelmath, Sims, 2010, p. 48; Мадиева, Уматова, 2020, с. 49).

В отличие от этого, в Национальном корпусе русского языка (НКРЯ) морфологический уровень разметки не является автономным и не исчерпывает аналитических задач. Флективная структура русского языка, отличающаяся синкретизмом грамматических значений и вариативностью синтаксических связей, требует многоуровневой интеграции морфологии и синтаксиса. В результате в НКРЯ реализована разветвлённая система морфосинтаксической и семантической аннотации (Богданов, Плунгян, Сичинава, 2005; Ляшевская, 2022). Данное различие подчёркивает, что в флективных языках многоуровневые схемы анализа являются не дополнительной опцией, а необходимостью для адекватного представления грамматических отношений.

Морфологическая аннотация как системообразующий уровень. Особый интерес представляет анализ принципов морфологической разметки. В корпусах казахского и турецкого языков морфемная сегментация выступает ключевым инструментом представления грамматической информации. Каждая словоформа декомпозируется на основу и последовательность аффиксов, что обеспечивает возможность точного поиска по морфологическим параметрам. Данный подход отражает структурную прозрачность агглютинативных систем и согласуется с положениями по теории агглютинативной морфологии (Haspelmath, Sims, 2010).

В Турецком национальном корпусе (Turkish National Corpus, TNC), помимо классического описания архитектуры, в последние годы были проведены значимые изменения, направленные на повышение качества аннотации и эффективности аналитического доступа к данным. Хотя основной корпус остаётся ориентированным на текстовый срез 1990–2009 гг., современные технологические улучшения включают оптимизацию поисковых механизмов, расширение морфологических фильтров и адаптацию аннотационных слоёв в соответствии с современными потребностями исследовательского сообщества. Эти изменения повышают аналитические возможности корпуса в задачах морфологической сегментации и лемматизации, делая его более сопоставимым по функциональности с аналогичными агглютинативными ресурсами (см. официальную документацию TNC).

В корпусе казахского языка аналогично делается ставка на морфологическую детализацию как фундаментальный элемент, обеспечиваю-

щий основу для построения последующих уровней анализа. Следует отметить, что улучшения в TNC происходили в диалоге с международным сообществом лингвистов и разработчиков, что создает дополнительные возможности для сравнительных исследований.

В НКРЯ морфологическая аннотация также присутствует, однако её роль в некоторых случаях носит вспомогательный характер по отношению к синтаксическому уровню. Это обусловлено тем, что флективное словоизменение в русском языке часто требует контекстной интерпретации, которую невозможно дать, опираясь исключительно на морфологические признаки.

Таким образом, результаты анализа подтверждают, что в агглютинативных языках морфология не только декларируется как ключевой уровень разметки, но и реализуется в корпусной архитектуре как системообразующая опция, в то время как в флективных языках важность синтаксической и семантической аннотации возрастает.

Синтаксическая и семантическая разметка. Сопоставление уровней аннотации показало, что НКРЯ обладает более глубокой синтаксической и семантической разметкой, чем корпуса агглютинативных языков. Развитая система зависимостного анализа и маркирования синтаксических ролей позволяет моделировать сложные грамматические отношения и выявлять синтаксические конструкции, характерные для русского языка. Эти особенности отражают традицию изучения сложных синтаксических единиц в славянских языках и подтверждаются данными исследований (McEnergy, Hardie, 2012).

В TNC синтаксическая аннотация также присутствует, однако её степень формализации остаётся более умеренной. Основное внимание уделяется морфологическим признакам, что объясняется приоритетами, связанными с морфемной структурой турецкого языка. При этом корпус активно интегрируется с ресурсами типа Universal Dependencies и treebank-сегментами (например, BOUN Treebank), что позволяет расширить аналитическое поле для задач синтаксического анализа и сопоставительного изучения языковых структур на межъязыковом уровне.

В казахском корпусе синтаксическая аннотация присутствует на базовом уровне, и её развитие осуществляется по мере наращивания текстовой базы и расширения программных интерфейсов. Современные исследования отмечают, что построение синтаксического слоя для

агглютинативных языков представляет собой сложную задачу, требующую адаптации традиционных моделей под специфику морфемной организации (специальные публикации по корпусной лингвистике агглютинативных языков).

Жанровая структура и репрезентативность. Анализ жанровой структуры корпусов показывает, что НКРЯ обладает выраженной диахронической глубиной, охватывая тексты различных исторических периодов и функциональных стилей. Наличие исторических подкорпусов обеспечивает возможность исследования динамики языковых изменений и сопоставления лингвистических явлений в разных эпохах.

В корпусах казахского и турецкого языков акцент смещён в сторону синхронного представления современного языка. При этом турецкий корпус активно использует дополнительные репрезентативные массивы, такие как большие веб-корпусы (например, trTenTen, trWaC) и парламентские корпуса (ParlaMint/Цифровой ParlaMint для турецкого языка), что даёт возможность сопоставлять традиционные корпусные данные с масштабными веб-корпусными частотностями. Эти ресурсы используются исследовательским сообществом начиная с 2020-2022 гг., что существенно расширяет аналитический потенциал корпуса (обзорные исследования по турецкому NLP).

Указанные жанровые различия отражают национальные предпочтения и исследовательские приоритеты. В контексте сравнительного анализа это позволяет учесть не только структурные параметры, но и распределение текстовых стилей — что особенно важно при сопоставлении межъязыковых явлений и типологических характеристик.

Сопоставление с предыдущими исследованиями. Сопоставление полученных результатов с ранее опубликованными исследованиями позволяет уточнить и конкретизировать научные выводы. В анализах, посвящённых корпусной архитектуре русского языка, подчеркивается многоуровневость аннотации и сложность синтаксического описания (Богданов, Плунгян, Сичинава, 2005; Ляшевская, 2022), однако сравнительные аспекты с корпусами агглютинативных языков остаются частично разработанными.

Со стороны исследований Турецкого национального корпуса наблюдается рост аналитической активности: начиная с 2020 г. публикуются кластерные обзоры ресурсов, сопоставляющие TNC с web-корпусами и treebank-структурами

в контексте NLP-задач (обозреваемые статьи 2021-2023). Эти публикации подчёркивают, что в практических задачах морфологическая аннотация TNC уже используется в сочетании с масштабными статистическими моделями, что расширяет практический потенциал корпуса.

В исследованиях Национального корпуса казахского языка отмечается, что детальный морфологический анализ является фундаментальным для правильной лемматизации и построения дальнейших аннотационных слоёв (Мадиева, Уматова, 2020), однако вопрос о синтаксической глубине предстоит развивать и интегрировать с международными стандартами (например, Universal Dependencies).

Настоящее исследование дополняет указанные подходы, формируя параметрическую модель сопоставления, в рамках которой архитектура национального корпуса рассматривается как типологически детерминированная цифровая система. Это позволяет не только формализовать выводы о влиянии морфологии, но и учитывать технологическую эволюцию ресурсов, такую как развитие TNC за последние годы.

На основе анализа и сопоставления корпусов можно сформулировать следующие научные положения:

- архитектура национального корпуса определяется типологическими характеристиками языка;
- в агглютинативных языках морфологическая аннотация играет ключевую системообразующую роль;
- во флективных языках приоритет смещён к многоуровневому морфосинтаксическому и семантическому описанию;
- технологическая эволюция ресурсов (в том числе TNC) расширяет аналитический потенциал корпуса и способствует интеграции корпусов с глобальными стандартами;
- жанровая структура корпуса отражает исследовательские и институциональные приоритеты национальной научной школы;
- типологически ориентированное сопоставление корпусов создаёт условия для более универсальных моделей цифрового описания языков.

Проведённый сопоставительный анализ позволяет дополнительно рассмотреть архитектуру национальных корпусов не только как совокупность технических решений, но и как отражение глубинных структурных свойств языка. В агглютинативных языках, таких как казахский

и турецкий, морфемная прозрачность и регулярность словообразовательных моделей обуславливают необходимость детализированной морфологической сегментации. Последовательное присоединение аффиксов формирует линейную морфемную структуру слова, что делает морфологический уровень аннотации ключевым системообразующим компонентом корпусной архитектуры. В этих условиях морфологическая разметка выступает не вспомогательным инструментом поиска, а базовым механизмом цифрового представления языковой системы.

В отличие от этого, во флективном русском языке синкретизм грамматических значений и вариативность словоизменительных парадигм формируют более сложную систему соотнесения формы и функции. Это обуславливает повышенное внимание к синтаксической аннотации и интеграции морфологического и синтаксиче-

ского уровней разметки. В модернизированной версии Национального корпуса русского языка 2.0 наблюдается тенденция к расширению зависимости разметки и усложнению поисковых параметров, что отражает необходимость моделирования синтаксических отношений в условиях флективной структуры языка.

Таким образом, выявляется закономерная корреляция между морфологическим типом языка и глубиной аннотационной стратификации корпуса. В агглютинативных системах приоритетным становится морфологический слой, обеспечивающий точную сегментацию и лемматизацию, тогда как во флективных языках особое значение приобретает синтаксическая и морфосинтаксическая аннотация. Это подтверждает положение о типологической детерминированности архитектуры национально-го корпуса.

Таблица 1

Типологическая корреляция архитектуры национальных корпусов

Параметр	Агглютинативные языки (казахский, турецкий)	Флективный язык (русский)
Морфологическая сегментация	Высокая степень морфемной прозрачности	Синкретизм грамматических форм
Приоритет аннотации	Морфологический уровень	Морфосинтаксический уровень
Лемматизация	Регулярная, аффиксальная	Парадигматическая, вариативная
Синтаксическая разметка	Дополнительная	Системообразующая
Архитектурная доминанта	Морфемная структура	Синтаксические зависимости

Примечание: таблица составлена авторами.

Представленные данные демонстрируют системную взаимосвязь между морфологическим типом языка и конфигурацией аннотационных уровней национального корпуса.

Вместе с тем анализ показывает, что корпусная архитектура формируется под воздействием не только структурных свойств языка, но и исследовательских традиций национальной научной школы. Казахская корпусная практика ориентирована на систематизацию морфологической информации и нормативный аспект языка, что отражает специфику развития казахской лингвистической традиции. В русской корпусной практике приоритет отдается расширению аналитического инструментария и многоуровневой аннотации. Турецкий национальный корпус

демонстрирует интеграцию традиционного корпусного подхода с современными цифровыми стандартами разметки.

Полученные результаты позволяют утверждать, что национальный корпус представляет собой не универсальную техническую платформу, а типологически обусловленную цифровую модель языковой системы. Следовательно, проектирование корпусных ресурсов должно учитывать морфологическую организацию языка как базовый фактор структурирования аннотационных уровней и поисковых механизмов. Игнорирование типологических особенностей может привести к формированию избыточных или, напротив, недостаточных аннотационных слоев, что снижает аналитическую эффективность корпуса.

Таким образом, проведённый анализ подтверждает рабочую гипотезу исследования и демонстрирует необходимость разработки типологически ориентированных принципов корпусного проектирования.

Заключение

Настоящее исследование было направлено на выявление зависимости архитектуры национального корпуса от морфологического типа языка на материале Национального корпуса казахского языка, Национального корпуса русского языка и Turkish National Corpus. В качестве методологической основы использован типологически ориентированный параметрический подход, позволивший сопоставить корпуса по совокупности структурных и функциональных критериев: глубине морфологической сегментации, степени формализации морфосинтаксической аннотации, жанровой конфигурации, диахроническому охвату и особенностям поискового инструментария.

Полученные результаты подтверждают выдвинутую гипотезу о типологической детерминированности корпусной архитектуры. Установлено, что в агглютинативных языках (казахский, турецкий) морфологическая аннотация выполняет системообразующую функцию и выступает базовым уровнем цифрового представления языковой структуры. Последовательная морфемная сегментация и прозрачность словообразовательной модели предопределяют приоритет морфологического слоя разметки в корпусной архитектуре.

В флективном языке (русский) приоритет смещается к интеграции морфологического и синтаксического анализа. Синкретизм грамматических форм и вариативность синтаксических связей обуславливают необходимость многоуровневой морфосинтаксической аннотации, что отражено в структуре НКРЯ, особенно в его модернизированной версии 2.0. Таким образом, корпусная архитектура в данном случае выступает не только средством фиксации языковых единиц, но и инструментом моделирования сложных грамматических отношений.

Сопоставительный анализ также показал, что конфигурация корпуса определяется не только внутренними структурными параметрами языка, но и исследовательскими приоритетами национальной научной школы. В казахстанской

лингвистической традиции особое внимание уделяется морфологической системности и кодификации нормы, что отражается в стратегии развития Национального корпуса казахского языка. В русской корпусной практике акцент сделан на расширении аннотационных слоёв и аналитических возможностей ресурса. Турецкая корпусная инфраструктура демонстрирует интеграцию национального корпуса с современными цифровыми ресурсами и зависимостными моделями разметки.

Научная новизна исследования заключается в разработке типологически ориентированной модели сопоставления национальных корпусов, в рамках которой корпус рассматривается как цифровая репрезентация морфологической организации языка. В отличие от описательных исследований отдельных проектов, предложенный подход позволяет выявить системную взаимосвязь между структурой языка и архитектурой корпусного ресурса.

Практическая значимость результатов заключается в возможности использования разработанной параметрической схемы при проектировании и модернизации национальных корпусов, а также при разработке аннотационных стандартов для языков различного морфологического типа. Полученные выводы могут быть применены в области автоматической обработки естественного языка, цифровой лексикографии и образовательных технологий.

Перспективы дальнейших исследований связаны с расширением типологического спектра анализируемых языков, включением полисинтетических и аналитических систем, а также с проведением количественного анализа аннотационных моделей на основе больших массивов корпусных данных. Это позволит уточнить степень универсальности выявленных закономерностей и углубить понимание взаимосвязи между типологией языка и его цифровым моделированием.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

Вклад авторов

Аляхимет Айдана Фабитқызы – разработка концепции исследования, сбор и систематизация материалов, проведение сравнительного анализа национальных языковых корпусов, ин-

терпретация результатов и подготовка основного текста статьи.

Уматова Жанна Максutowна – научно-методологическое сопровождение исследования, формирование теоретической базы, участие в

интерпретации результатов, научное редактирование текста и подготовка заключительной части статьи.

Все авторы ознакомились с окончательной версией статьи и одобрили её к публикации.

Литература

- Аханов, К. (2002). *Тіл білімінің негіздері*. Алматы: Санат.
- Жұбанов, Қ. (1999). *Қазақ тілі жөніндегі зерттеулер*. Алматы: Ғылым.
- Қайдар, Ә. (1998). *Қазақ тілінің өзекті мәселелері*. Алматы: Ана тілі.
- Уәли, Н. (2012). *Қазақ әдеби тілінің нормалары*. Алматы.
- Құрманбайұлы, Ш. (2005). *Терминқор қалыптастыру көздері мен терминжасам тәсілдері*. Алматы: Сөздік-Словарь.
- Comrie, B. (1989). *Language universals and linguistic typology*. Chicago: University of Chicago Press.
- Haspelmath, M., & Sims, A. (2010). *Understanding morphology*. London: Routledge.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies*. London: Routledge.
- McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
- Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.
- Gries, S. T. (2019). *Quantitative corpus linguistics with R*. London: Routledge.
- Ide, N., & Pustejovsky, J. (2017). *Handbook of linguistic annotation*. Dordrecht: Springer.
- McEnery, T., & Brezina, V. (2022). *Fundamental principles of corpus linguistics*. Cambridge: Cambridge University Press.
- Nivre, J., et al. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. *Proceedings of LREC*.
- Богданов, С. И., Плунгян, В. А., & Сичинава, Д. В. (2005). Национальный корпус русского языка: принципы построения и перспективы развития. *Вопросы языкознания*, 3, 3–14.
- Aksan, Y., Aksan, M., Mersinli, Ü., & Demirhan, U. (2012). Construction of the Turkish National Corpus (TNC). *Proceedings of LREC*, 3223–3227.
- Мадиева, Г. Б., & Уматова, Ж. М. (2020). Национальный корпус казахского языка: принципы формирования и перспективы развития. *Вестник КазНУ. Серия филологическая*, 2(176), 45–55.
- Фазылжанова, А. Б. (2021). Цифровизация казахского языка и корпусные технологии. *Қазақ тілі мен әдебиеті*, 3, 12–19.
- Ляшевская, О. Н. (2022). Национальный корпус русского языка: современное состояние и перспективы развития. *Компьютерная лингвистика и интеллектуальные технологии*, 21, 40–48.
- Савчук, С. О., Архангельский, Т. А., Bonch Osmolovskaya, A. A., Дони́на, О. В., Кузнецова, Ю. Н., Ляшевская, О. Н., Орехов, Б. В., & Подрядчикова, М. В. (2024). Национальный корпус русского языка 2.0: новые возможности и перспективы развития. *Вопросы языкознания*, 2, 7–34.

References

- Akhanov, K. (2002). *Til biliminin negizderi* [Foundations of linguistics]. Almaty: Sanat. (In Kazakh)
- Aksan, Y., Aksan, M., Mersinli, Ü., & Demirhan, U. (2012). Construction of the Turkish National Corpus (TNC). *Proceedings of LREC*, 3223–3227.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
- Bogdanov, S. I., Plungyan, V. A., & Sichinava, D. V. (2005). Natsional'nyi korpus russkogo yazyka: printsipy postroeniya i perspektivy razvitiya [The Russian National Corpus: Principles of design and development prospects]. *Voprosy Yazykoznaniya*, 3, 3–14. (In Russian)
- Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.
- Comrie, B. (1989). *Language universals and linguistic typology*. Chicago: University of Chicago Press.
- Fazylzhanova, A. B. (2021). Tsifrovizatsiya kazakhskogo yazyka i korpusnye tekhnologii [Digitalization of the Kazakh language and corpus technologies]. *Qazaq Tili men Adebieti*, 3, 12–19. (In Russian)
- Gries, S. T. (2019). *Quantitative corpus linguistics with R*. London: Routledge.
- Haspelmath, M., & Sims, A. (2010). *Understanding morphology*. London: Routledge.
- Ide, N., & Pustejovsky, J. (2017). *Handbook of linguistic annotation*. Dordrecht: Springer.
- Kaidar, A. (1998). *Qazaq tilinin ozekti maseleleri* [Topical issues of the Kazakh language]. Almaty: Ana tili. (In Kazakh)
- Kurmanbaiuly, Sh. (2005). *Terminqor qalyptastyru kozderi men terminzhasam tasilderi* [Sources of termbase formation and methods of term formation]. Almaty: Sozdik-Slovar'. (In Kazakh)

Lyashevskaya, O. N. (2022). Natsional'nyi korpus russkogo yazyka: sovremennoe sostoyanie i perspektivy razvitiya [The Russian National Corpus: Current state and development prospects]. *Komp'yuternaia Lingvistika i Intellekturnye Tekhnologii*, 21, 40–48. (In Russian)

Madieva, G. B., & Umatova, Zh. M. (2020). Natsional'nyi korpus kazakhskogo yazyka: printsipy formirovaniya i perspektivy razvitiya [The Kazakh National Corpus: Principles of formation and development prospects]. *Vestnik KazNU. Seriya Filologicheskaya*, 2(176), 45–55. (In Russian)

McEnery, T., & Brezina, V. (2022). *Fundamental principles of corpus linguistics*. Cambridge: Cambridge University Press.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies*. London: Routledge.

Nivre, J., et al. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. *Proceedings of LREC*.

Savchuk, S. O., Arkhangel'skii, T. A., Bonch Osmolovskaya, A. A., Donina, O. V., Kuznetsova, Yu. N., Lyashevskaya, O. N., Orekhov, B. V., & Podriadchikova, M. V. (2024). Natsional'nyi korpus russkogo yazyka 2.0: novye vozmozhnosti i perspektivy razvitiya [The Russian National Corpus 2.0: New opportunities and development prospects]. *Voprosy Yazykoznanii*, 2, 7–34. (In Russian)

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Uali, N. (2012). *Qazaq adebi tilinin normalary* [Norms of the Kazakh literary language]. Almaty. (In Kazakh)

Zhubanov, Q. (1999). *Qazaq tili zhonindegi zertteuler* [Studies on the Kazakh language]. Almaty: Gylm. (In Kazakh)

Сведения об авторах:

Аляхимет Айдана Ғабитқызы (автор-корреспондент) – PhD докторант, Казахский национальный университет имени аль-Фараби (Алматы, Казахстан, e-mail: alyakhimet.aidana@gmail.com).

Уматова Жанна Максutowна – кандидат филологических наук, доцент, Синьцзянский университет (Урумчи, Куйтайская Народная Республика, e-mail: zh.umatova@mail.ru).

Information about the authors:

Alyakhimet Aidana (corresponding author) – PhD student, al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: alyakhimet.aidana@gmail.com).

Umatova Zhanna – Candidate of Philological Sciences, Associate Professor, Xinjiang University (Ürümqi, People's Republic of China, e-mail: zh.umatova@mail.ru).

Авторлар туралы мәлімет:

Аляхимет Айдана Ғабитқызы (корреспонденттік автор) – PhD докторант, Әл-Фараби атындағы Қазақ ұлттық университеті (Алматы, Қазақстан, e-mail: alyakhimet.aidana@gmail.com);

Уматова Жанна Максutowна – филология ғылымдарының кандидаты, доцент, Шыңжаң университеті (Үрімші, Қытай Халық Республикасы, e-mail: zh.umatova@mail.ru).

Поступила: 5 января 2026 г.

Принята: 9 июня 2026 г.