MACROSOCIOLINGUISTICS AND MINORITY LANGUAGES 2025;3(2):97–116
ISSN 2949-5997 (Online) https://macrosociolingusictics.ru/MML/

ЯЗЫКИ НАРОДОВ РОССИЙСКОЙ ФЕДЕРАЦИИ: ЦИФРОВЫЕ ИНСТРУМЕНТЫ ДОКУМЕНТИРОВАНИЯ И МЕДИАДОСТУПНОСТЬ
THE LANGUAGES OF THE PEOPLES OF RUSSIAN FEDERATION: DIGITAL DOCUMENTATION TOOLS AND MEDIA ACCESSIBILITY

DOI: 10.22363/2949-5997-2025-3-2-97-116
EDN HARJOQ

Научная статья / Research article

Письменный корпус татарского языка: структура, состав и возможности использования

Шахдар Нурдиновна КЕНЕШБЕКОВА

Российский университет дружбы народов, *Москва, Российская Федерация*
Российский новый университет, *Москва, Российская Федерация*
✉ keneshbekova.sh@yandex.ru

Аннотация. Исследование носит аналитический характер и посвящено изучению и описанию структуры, состава и функциональных возможностей Письменного корпуса татарского языка (Корпус) — одного из наиболее масштабных цифровых ресурсов для тюркских языков Российской Федерации. Рассмотрены этапы создания Корпуса, методология сбора и аннотирования текстов, включая метаразметку и морфологическую разметку с использованием системы *Apertium*. Особое внимание уделено прикладным аспектам: интеграции Корпуса в системы синтеза и распознавания речи, разработке различных лингвистических сервисов, а также его применению в образовательных и научных проектах. Проанализированы принципы устранения дублирующихся текстов и предложены перспективы дальнейшего развития, включая расширение жанрового разнообразия и внедрение международных стандартов аннотирования. Материалом исследования послужили сам Корпус и публикации, описывающие этапы его создания и применения. Методология представлена комплексом эмпирических методов и приемов, как наблюдение, анализ, описание, тестирование (функциональных возможностей корпуса и др.), а также графический метод для визуализации изучаемого материала. Отмечена научная и культурная значимость Письменного корпуса татарского языка в контексте цифровизации языков народов России, что соответствует задачам Международного десятилетия языков

© Кенешбекова Ш.Н., 2025

This work is licensed under a Creative Commons Attribution 4.0 International License
https://creativecommons.org/licenses/by-nc/4.0/legalcode

коренных народов (2022–2032), инициированного Генеральной Ассамблей ООН и координируемого Организацией Объединенных Наций по вопросам образования, науки и культуры.

Ключевые слова: корпусная лингвистика, языки народов России, киберэтнография, цифровая гуманитаристика

Заявления о конфликте интересов. Автор заявляет об отсутствии конфликта интересов.

История статьи: получена 30 сентября 2025; принята в печать 14 октября 2025.

Для цитирования: *Кенешбекова Ш.Н.* Письменный корпус татарского языка: структура, состав и возможности использования // Macrosociolinguistics and Minority Languages. 2025. Т. 3. № 2. С. 97–116. https://doi.org/10.22363/2949-5997-2025-3-2-97-116 EDN: HARJOQ

Corpus of written Tatar: structure, composition and applications

Shakhdar N. KENESHBEKOVA

RUDN University, *Moscow, Russian Federation*
Russian New University, *Moscow, Russian Federation*
✉ keneshbekova.sh@yandex.ru

Abstract. The proposed study is analytical in nature and is devoted to examining and describing the structure, composition, and functional capabilities of the Written Corpus of the Tatar language, which represents one of the largest digital resources for the Turkic languages of the Russian Federation. The study discusses the stages of corpus creation and the methodology of text collection and annotation, including metadata annotation and morphological annotation using the *Apertium* system. Special attention is paid to applied aspects, such as the integration of the corpus into speech synthesis and speech recognition systems, the development of various linguistic services and its use in educational and research projects. The principles of eliminating duplicate texts are analyzed and prospects for further development are proposed, including the expansion of genre diversity and the introduction of international annotation standards. The material for this study comprises the corpus itself and publications describing the stages of its creation and application. The methodology is presented as a set of empirical methods and techniques, including observation, analysis, description, and testing (of the functional capabilities of the corpus, etc.), as well as the graphical method for visualizing the material under study. The study highlights the scientific and cultural significance of the Written Corpus of the Tatar Language in the context of the digitalization of the languages of the peoples of Russia, which corresponds to the objectives of the International Decade of Indigenous Languages (2022–2032), initiated by the United Nations General Assembly and coordinated by the United Nations Educational, Scientific and Cultural Organization.

Key words: corpus linguistics, languages of Russian Federation, Tatar language, cyberethnography, digital humanities
Conflict of interest. The author declares no conflicts of interest.

Article history: received 30 September 2025; accepted 14 October 2025.

For citation: Keneshbekova, Sh.N. (2025). Corpus of written Tatar: Structure, composition and applications. *Macrosociolinguistics and Minority Languages, 3*(2), 97–116. (In Russ.). https://doi.org/110.22363/2949-5997-2025-3-2-97-116 EDN: HARJOQ

Введение

Разработка лингвистических корпусов для языков народов России занимает важное место в современной языковой политике и языковом планировании страны. Государство поддерживает подобные лингвистические инициативы, что подтверждается, например, Законом РФ «О языках народов Российской Федерации»[1] с изменениями от 13.06.2023 № 253-ФЗ, а также на Распоряжение Правительства Российской Федерации от 12.06.2024 № 1481-р[2]. Первый признает языки народов России (ЯНР) национальным достоянием государства, определяя их место в системе образования, средствах массовой информации, производстве и других областях деятельности. Второй — определяет языковое многообразие РФ совокупностью «функционирующих языков, диалектов и говоров народов Российской Федерации»[3], утверждая, что республики страны «наделены правом устанавливать свои государственные языки, функционирующие наряду с государственным»[4]. Регламентируя области функционирования ЯНР, данное Распоряжение обращает внимание на «содействие изданию литературы на языках народов Российской Федерации, <…> переводческой деятельности, финансирование научных исследований в области сохранения, изучения и развития языков народов Российской Федерации, создание условий для распространения через средства массовой информации сообщений и материалов на языках народов Российской Федерации, подготовку специалистов в указанной области, совершенствование системы образования в целях развития языков народов Российской Федерации»[5]. В таком контексте, с точки зрения государственной поддержки, разработка, развитие и научно-образовательная популяризация корпусов ЯНР видится актуальным и востребованным направлением деятельности современной фундаментальной и прикладной лингвистики.

[1] Закон РФ «О языках народов Российской Федерации» от 25.10.1001 № 1807–1 с изменениями от 13.06.2023 № 253-ФЗ. URL: https://www.consultant.ru/document/cons_doc_LAW_15524/ (дата обращения: 15.08.2025)
[2] Распоряжение Правительства Российской Федерации от 12.06.2024 № 1481-р. URL: http://publication.pravo.gov.ru/document/0001202406140048?index=1 (дата обращения: 15.08.2025).
[3] Распоряжение Правительства Российской Федерации от 12.06.2024 № 1481-р. С. 3. URL: http://publication.pravo.gov.ru/document/0001202406140048?index=1 (дата обращения: 15.08.2025).
[4] Распоряжение Правительства Российской Федерации от 12.06.2024 № 1481-р. С. 4. URL: http://publication.pravo.gov.ru/document/0001202406140048?index=1 (дата обращения: 15.08.2025).
[5] Распоряжение Правительства Российской Федерации от 12.06.2024 № 1481-р. С. 6. URL: http://publication.pravo.gov.ru/document/0001202406140048?index=1 (дата обращения: 15.08.2025).

В собственно научной перспективе разработка корпусов для ЯНР встраивается в область интересов киберэтнографии в аспекте фиксации, упорядочивания и описания представленности того или иного языка в веб-среде. Эти задачи обусловлены большей «ориентированностью» киберэтнографии «на тексты <…>, а не на наблюдение и взаимодействие» с представителями изучаемого народа (Белоруссова, 2021: 127).

Корпусные технологии позволяют не только сохранять и изучать языковое разнообразие, но и открывают новые возможности для цифровой обработки языка, машинного перевода, автоматического анализа текстов и других прикладных задач (Алюнина, 2025: 102–103). Так разработка лингвистических корпусов, особенно параллельных (Алюнина, 2025: 103–104), соответствует вектору развития программ машинного перевода, которые с начала 2020-х гг. активно пополняются ЯНР: *Переводчик Яндекс*[6] (бурятский, коми, мокшанский, татарский, удмуртский, чувашский и др.), *Google Translate*[7] (бурятский, коми, осетинский, татарский и др.), *Microsoft Bing*[8] (башкирский, татарский и др.), *PROMT.One*[9] (татарский) и др. Особую значимость деятельность по разработке подобных веб-ресурсов приобретает в условиях глобализации, когда многие языки, и не только миноритарные, сталкиваются с угрозой цифровой маргинализации — вытеснения из электронного пространства «языками-первопроходцами», такими как английский, китайский, русский, арабский, испанский, итальянский, немецкий, французский, итальянский, о чем свидетельствуют статистические данные разных лет, начиная с конца 1990-х до начала 2020-х гг.[10, 11]

На сегодняшний день в России ведется работа по созданию корпусов для многих ЯНР, что способствует их документированию и интеграции в цифровую среду. Среди примеров можно назвать следующие проекты:

- Бурятский корпус[12] — проект Института монголоведения, буддологии и тибетологии Сибирского отделения РАН при участии специалистов Института востоковедения РАН, Лионского университета во Франции, Университета Гамбурга в Германии;

[6] Переводчик Яндекс. URL: https://translate.yandex.ru/translator/%D0 %90 %D0%BD%D0 %B3 %D0 %BB% D0 %B8 %D0 %B9 %D1 %81%D0 %BA%D0 %B8 %D0 %B9-%D0 %A0 %D1 %83 %D1 %81 %D1 % 81 %D0 %BA%D0 %B8 %D0 %B9 (дата обращения: 16.07.2025)
[7] Google Translate. URL: https://translate.google.com/?hl=ru&tab=TT&sl=ru&tl=en&op=translate (дата обращения: 16.07.2025).
[8] Microsoft Bing. URL: https://www.bing.com/translator?from=&to=ru&setlang=ru (дата обращения: 16.07.2025).
[9] PROMT.One. URL: https://www.translate.ru/%D0 %BF%D0 %B5 %D1 %80 %D0 %B5 %D0 %B2 %D0 %BE%D0 %B 4 (дата обращения: 16.07.2025).
[10] Usage statistics of content languages for websites. URL: https://w3techs.com/technologies/overview/content_language (дата обращения: 20.07.2025).
[11] *Pimienta D., Prado D., Blanco A.* Twelve years of measuring linguistic diversity in the Internet: balance and perspectives. Paris : UNESCO Publications for the World Summit on the Information Society, 2009. 58 p. URL: https://unesdoc.unesco.org/ark:/48223/pf0000187016 (дата обращения: 20.07.2025).
[12] Бурятский корпус. URL: http://web-corpora.net/BuryatCorpus/search/?interface_language=ru (дата обращения: 16.07.2025).

• Открытый корпус вепсского и карельского языков[13] — разрабатывается сотрудниками Карельского научного центра РАН (Родионова, Пеллинен, 2024);

• Корпус мансийского языка[14] — проект Института языкознания РАН при участии Института русского языка им. В.В. Виноградова, МГУ им. М.В. Ломоносова, Новгородского государственного университета им. Ярослава Мудрого, Обско-угорского института прикладных исследований и разработок);

• Корпус удмуртского языка[15] (основной корпус, корпус соцсетей, звуковой корпус) — проект Школы лингвистики Научно-исследовательского университета «Высшая школа экономики» (далее — ВШЭ);

• Электронный корпус чувашского языка[16] — разрабатывается Чувашским государственным институтом гуманитарных наук;

• Татарский национальный корпус «Туган тел»[17] — проект, реализуемый Научно-исследовательским институтом «Прикладная семиотика» Академии наук Республики Татарстан, Казанским (Приволжским) федеральным университетом и ВШЭ;

• Письменный корпус татарского языка[18] — проект Университета Иннополис в России, Университета Квебека в Канаде, Штутгартского университета в Германии, Университета Турку в Финляндии и др.

Среди существующих сегодня корпусных инициатив Письменный корпус татарского языка (ПКТЯ) является одним из наиболее развитых корпусов тюркских языков в России. Мы поставили цель: комплексно рассмотреть структуру, состав и функциональные возможности ПКТЯ и проанализировать его роль в современных лингвистических и прикладных исследованиях. Материалом для анализа служит ПКТЯ, включающий тексты различных жанров и периодов, а также его техническая и лингвистическая разметка.

Структура статьи воспроизводит последовательный анализ следующих аспектов ПКТЯ: 1) историю разработки и состав корпуса; 2) функциональные возможности, поисковый и аналитический потенциал; 3) применение в фундаментальных и прикладных исследованиях; 4) а также перспективы развития и использования функционала и наполнения ПКТЯ, что позволяет оценить его вклад в поддержку татарского языка, а также роль корпуса в развитии современных проектов в области цифровой гуманитаристики (Digital Humanities).

[13] Открытый корпус вепсского и карельского языков. URL: http://dictorpus.krc.karelia.ru/ru (дата обращения: 16.07.2025).
[14] Аннотированный корпус мансийских текстов. URL: https://mansi.pro/corpus/ (дата обращения: 16.07.2025).
[15] Корпуса удмуртского языка. URL: http://web-corpora.net/UdmurtCorpus/search/?interface_language=ru (дата обращения: 16.07.2025).
[16] Электронный корпус чувашского языка. URL: https://chuvkorpus.ru/ (дата обращения: 17.07.2025).
[17] Татарский национальный корпус «Туган тел». URL: http://web-corpora.net/TatarCorpus/search/?interface_language=ru (дата обращения: 16.07.2025).
[18] Письменный корпус татарского языка. URL: https://www.corpus.tatar/ (дата обращения: 16.07.2025).

История разработки Письменного корпуса татарского языка и его состав

Инициатива создания ПКТЯ принадлежит коллективу исследователей, включающему специалистов из России, Финляндии, Германии и Канады, которые начали работу над проектом в 2010 г. Партнерами проекта стали редакция научно-информационного журнала «Фән һәм Тел» — «Наука и язык»[19], Республиканская специальная библиотека для слепых и слабовидящих в Татарстане[20], Институт языка, литературы и искусства имени Г. Ибрагимова Академии наук Республики Татарстан[21], разработчики корпусного менеджера *Sketch Engine*[22] и др. Таким образом, команда ПКТЯ представлена филологами, лингвистами, инженерами, специалистами сферы информационной и медиадоступности, что обеспечивало междисциплинарный подход к разработке рассматриваемого лингвистического ресурса.

Первый этап создания ПКТЯ включал проектирование его архитектуры, разработку поискового механизма и определение принципов выборки текстов. К марту 2012 г. была завершена базовая версия корпуса, включающая веб-интерфейс и систему поиска. Первая публичная версия корпуса, содержащая 45 млн словоупотреблений из 60 источников различных жанров и стилей, стала доступна в начале 2012 г. (Сайхунов, Хусаинов, Ибрагимов, 2018: 314).

В 2014 г. выпущена вторая версия корпуса, объем данных которого увеличился до 116 млн словоупотреблений из 2750 источников (Ибрагимов, Сайхунов, 2014: 261). Под версией или релизом корпуса в данном случае понимается обновление его текстовой базы. Для расширения функциональных возможностей ПКТЯ в период между 2024 и 2018 гг. в корпус была внедрена морфологическая разметка, в основу которой легла «система автоматической грамматической аннотации» (Сайхунов, Хусаинов, Ибрагимов, 2018: 314), разработанная *Apertium*[23] и адаптированная к большому количеству языков, в т.ч. к татарскому. Жанрово-стилистический состав корпуса в его второй итерации приведен на рис. 1.

После релиза второй версии ПКТЯ корпус был дополнен возможностью проверки правописания на татарском языке (2017 г.) и функцией синтеза татарской речи (2015 г.), в разработке которой участвовали специалисты Республиканской специальной библиотеки для слепых и слабовидящих[24].

[19] «Фән һәм тел» — «Наука и язык»: научно-информационный журнал. URL: https://tatarica.org/ru/razdely/sredstva-massovoj-informacii/periodicheskie-izdaniya/fehn-hehm-tel (дата обращения: 18.07.2025).
[20] ГБУК Республики Татарстан «Республиканская специальная библиотека для слепых и слабовидящих имени Ш.К. Еникеева». URL: https://rsbsrt.ru/ (дата обращения: 18.07.2025).
[21] Институт языка, литературы и искусства имени Г. Ибрагимова Академии наук Республики Татарстан https://www.antat.ru/ru/iyli/ (дата обращения: 19.07.2025).
[22] Sketch Engine. URL: https://www.sketchengine.eu/ (дата обращения: 19.07.2025).
[23] Apertium. URL: https://www.apertium.org/index.rus.html#?dir=rus-bel&q= (дата обращения: 17.08.2025).
[24] ГБУК Республики Татарстан «Республиканская специальная библиотека для слепых и слабовидящих имени Ш.К. Еникеева». URL: https://rsbsrt.ru/ (дата обращения: 18.07.2025).

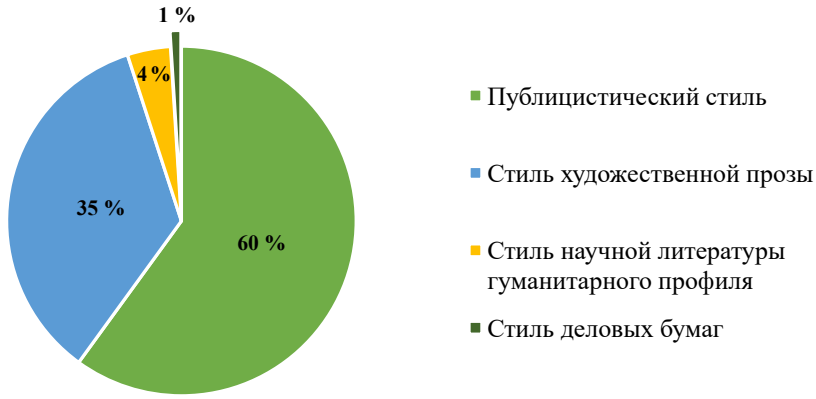

Рис. 1. Жанрово-стилевой состав ПКТЯ по состоянию на 2014 г.
Источник: выполнила Ш.Н. Кенешбекова по материалам (Ибрагимов, Сайхунов, 2014: 261).

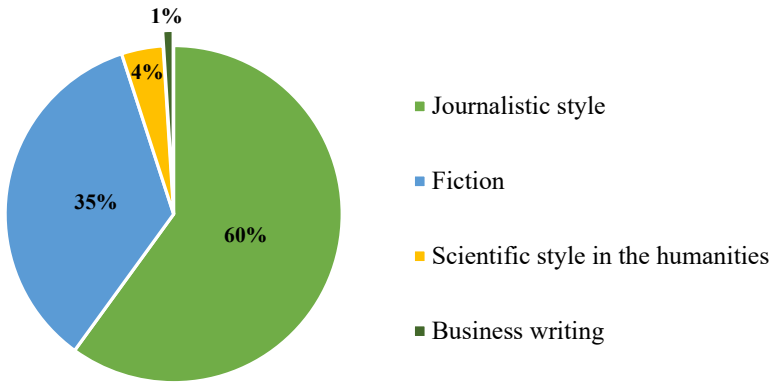

Fig. 1. Genre and style composition of the Corpus of Written Tatar as of 2014
Source: compiled by Sh.N. Keneshbekova after (Ibragimov, Saikhunov, 2014: 261).

Третья версия корпуса, выпущенная в конце 2018 г., почти втрое превысила объем предыдущей, — 356 млн слов (430 млн токенов) из 16 786 текстов (Сайхунов, Хусаинов, Ибрагимов, 2019: 548); перечень функциональных возможностей пополнился разделом Тезаурус.

Последнее, четвертое обновление ПКТЯ, состоявшееся в конце 2019 г., увеличило его объем до 500 млн словоупотреблений (17 000 источников) и привнесло в корпус обновленную систему проверки правописания для татарского языка, улучшенную морфологическую разметку. Позже, в 2022 г., у ПКТЯ появился специализированный раздел «Личные имена», включающий перечень татарских антропонимов — имен, отчеств и фамилий. Пример результата выполнения поискового запроса по татарскому личному имени в ПКТЯ приведен на изображении далее (рис. 2).

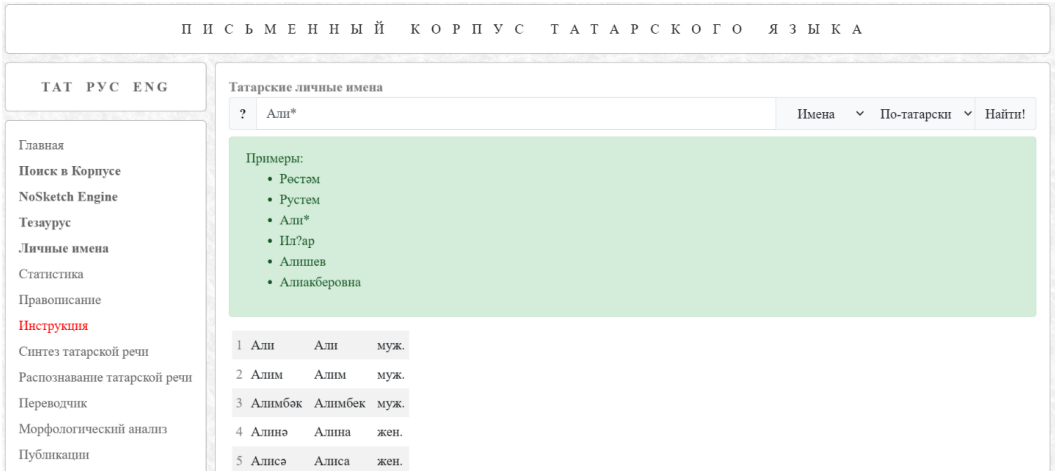

Рис. 2. Скриншот с результатом поиска татарского имени

Источник: Письменный корпус татарского языка. URL: https://www.corpus.tatar/index.php?of=search/names.php (дата обращения: 12.08.2025).

Fig. 2. Screenshot with the search result for a Tatar name

Source: Corpus of written Tatar. Retrieved 12 August 2025, from: https://www.corpus.tatar/index.php?of=search/names.php

Таким образом, основной целью разработки ПКТЯ стало создание инструментария для цифровой фиксации татарского языка и систематизации процесса его изучения. Также корпус может быть использован для преподавания татарского языка, разработки методических материалов, пополнения обучающих баз машинных переводчиков в языковой паре с татарским языком, автоматического синтеза и распознавания татарской речи, что делает вклад в развитие доступной среды. Помимо этого, корпус выполняет функцию важного ресурса для лингвистических исследований, предоставляя доступ к обширной и репрезентативной текстовой базе.

Процесс разработки ПКТЯ сопряжен с рядом технологических и организационных вызовов, которые включают отсутствие централизованного финансирования, необходимость адаптации программных решений под специфику татарского языка, а также привлечение экспертов в области корпусной лингвистики. Интерфейс корпуса доступен на трех языках (татарском, русском и английском), что способствует его интеграции в международное научное сообщество. Вместе с тем рост объема корпусов сопряжен с возрастающей сложностью в обеспечении уникальности содержащихся в них текстовых единиц. В процессе корпусной компиляции используются различные методологические подходы к устранению дублирования, основанные на лингвистическом анализе и алгоритмических решениях.

Функциональные возможности Письменного корпуса татарского языка

Письменный корпус татарского языка представляет собой универсальную цифровую платформу, предназначенную для комплексного анализа структуры и функционирования татарского языка. Корпус широко используется как в академических целях, так и для решения прикладных задач: разработка образовательных и методических материалов, лексикографических ресурсов разных типов и технологий автоматической обработки текста и речи, подготовка лингвистически аннотированных ресурсов для обучения и NLP-систем и др. (Кузнецов, 2023; Гатиатуллин и др., 2024; Galieva, Vavilova, Gafarova, 2017; Luutonen, Moisio, Daher, 2017).

Встроенные в корпус программные средства обеспечивают широкий спектр функциональных возможностей, охватывающих:

- поиск и количественный анализ лексических единиц (реализуется функцией Статистика);
- анализ лексической и грамматической сочетаемости, т.е. поиск типичных для языка коллокаций и коллигаций (реализуется, например, функциями n-грамм, KWIC — Key word in context — ключевое слово в контексте);
- поиск по морфологическим, семантическим и стилистическим признакам (реализуется с помощью настройки поискового запроса в интерфейсе ПКТЯ);
- проверку орфографии на татарском языке (функция Правописание);
- озвучивание письменного текста на татарском языке и транскрибирование татарской речи (функции Синтез татарской речи и Распознавание татарской речи).

Приведенные функции позволяют решать широкий спектр задач, а за их реализацией и достоверностью прогнозируемых результатов поиска стоит строгая система корпусного аннотирования или корпусной разметки, под которой понимается лингвистическая информация, «приписываемая всем единицам выбранного уровня: текст, предложение, словосочетание, словоформа» и др. (Алюнина, 2025: 96). В зависимости от степени обращения к программным ресурсам в аннотировании корпусных документов выделяют ручную и автоматическую разметку (Алюнина, 2025: 98).

Метаразметка и классификация собранных текстов

С технической точки зрения корпус создается с применением методов автоматической обработки текстов, включая морфологическую разметку и поисковые алгоритмы. Однако значительная часть работы, такая как верификация данных и категоризация текстов, выполняется вручную. Как было

показано на рис. 1, источниками для пополнения корпуса служат письменные тексты различного происхождения: художественная литература, публицистика, научные и официальные документы.

Метаразметка в составе ПКТЯ представляет собой систему аннотирования текстов с использованием описательных характеристик, которые обеспечивают упорядоченное хранение, систематизированный поиск и последующую аналитическую обработку данных.

В международной корпусной практике в качестве методологических ориентиров широко применяются стандарты *Text Encoding Initiative — Стандарт кодирования текстовых данных (TEI)*[25] и рекомендации *Expert Advisory Group on Language Engineering Standards — Экспертной группы по стандартизации структуры лингвистических ресурсов (EAGLES)*[26], на базе которых разработан *Corpus Encoding Standard — Перечень рекомендаций корпусной разметки (CES)*[27]. Эти стандарты обеспечивают единообразие в представлении метаданных. Как правило, они реализуются в формате XML, что способствует совместимости корпусных ресурсов и их интероперабельности на межплатформенном уровне (Сайхунов, Хусаинов, Ибрагимов, 2019: 549).

В отличие от названных инструментов, ПКТЯ использует собственную, адаптированную к его специфике систему метаразметки. Формально она реализована в виде так называемых plain text-структур, не прибегающих к полноформатной XML-разметке. Подобное решение обусловлено рядом практических причин: текстовые файлы в «плоском» формате легче обрабатываются базовыми утилитами UNIX-подобных систем (такими как awk, sed, grep, sort, cut, paste и др.)[28], обладают высокой степенью читаемости и не требуют специализированного программного обеспечения для начального этапа анализа (Сайхунов, Хусаинов, Ибрагимов, 2019: 550).

Несмотря на функциональность существующей модели, дальнейшая эволюция корпуса предполагает постепенное внедрение унифицированных международных стандартов метаразметки. Переход к XML-структурам и расширенным схемам аннотации позволит достичь следующих целей:

- интеграции корпуса в многоязычные исследовательские платформы;
- повышения точности и детализации лингвистических исследований;
- обеспечения прозрачного обмена данными между научными учреждениями и проектами.

На скриншоте далее представлен пример интерфейса, где демонстрируется структура метаразметки текста в ПКТЯ (рис. 3).

[25] Text Encoding Initiative. URL: https://tei-c.org/ (дата обращения: 23.08.2025).
[26] Expert Advisory Group on Language Engineering Standards. URL: https://cordis.europa.eu/project/id/LE34244 (дата обращения: 23.08.2025).
[27] Corpus Encoding Standard. URL: https://www.cs.vassar.edu/CES/ (дата обращения: 23.08.2025).
[28] 13 инструментов для обработки текста в командной оболочке // Хабр. URL: https://habr.com/ru/companies/itsumma/articles/492932/ (дата обращения: 18.08.2025).

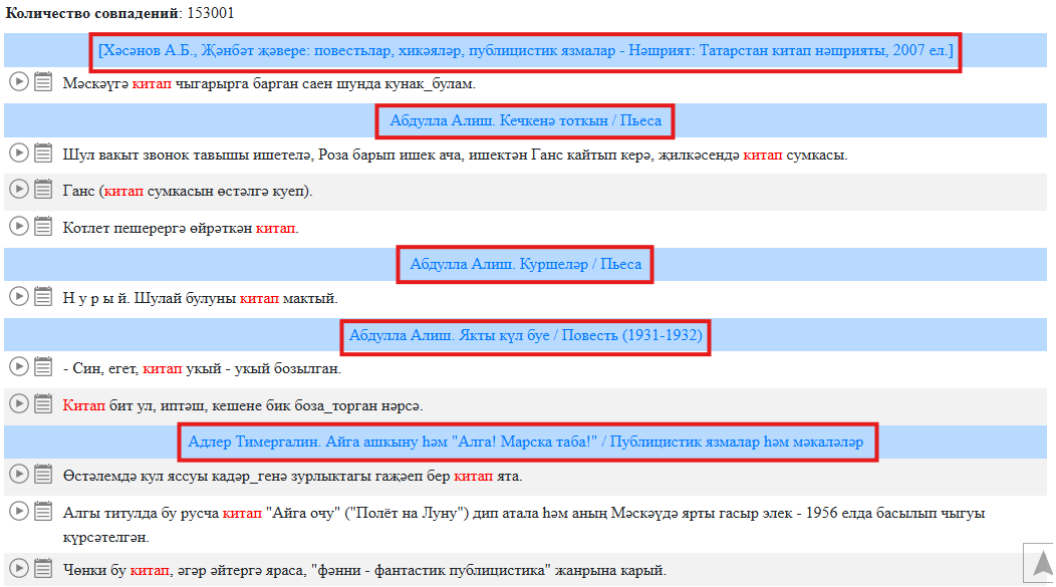

Рис. 3. Метаданные в Письменном корпусе татарского языка

Источник: Письменный корпус татарского языка. URL: https://www.corpus.tatar/ru (дата обращения: 12.08.2025).

Fig. 3. Metadata in the Corpus of Written Tatar

Source: Corpus of written Tatar. Retrieved 12 August 2025, from: https://www.corpus.tatar/index.php?of=search/names.php

Как показано на рис. 3, на текущем (по состоянию на август 2025 г.) этапе в корпус включены следующие параметры экстралингвистической метаразметки, которые обеспечивают базовую классификацию текстового материала (Сайхунов, Хусаинов, Ибрагимов, 2019: 550; Алюнина, 2025: 100):

- авторство — указание имени автора или авторского коллектива;
- название — заголовок произведения, статьи, книги, либо интернет-источника;
- хронологическая метка — сведения о времени создания или публикации текста;
- типологическая характеристика — сведения о стилистической принадлежности текста (художественный, публицистический, научный, официально-деловой, фольклорный и т.д.);
- жанровая принадлежность — сведения о жанровых характеристиках текста (роман, рассказ, статья, поэма, сказание и др.);
- источник — библиографическая ссылка на издание или цифровую платформу;
- URL-адрес — сведения об электронном источнике, которые указываются в случае включения в корпус текста с цифрового ресурса;

• дополнительные атрибуты — техническая или организационная информация (кодировка, версия разметки, ID источника, поставщик данных и пр.).

Морфологическая разметка

Под морфологической разметкой лингвистического корпуса понимается присвоение «каждой словоформе ее морфологических признаков» (Алюнина, 2025: 99). С утилитарной точки зрения этот тип разметки обеспечивает возможность поиска по морфологическим параметрам, например, поиск существительного мужского рода в форме множественного числа. В процессе морфологического аннотирования ПКТЯ применялся ресурс *Apertium*[29] — свободное программное обеспечение с полностью открытым исходным кодом, широко используемое в корпусной лингвистике благодаря своей модульности и адаптивности. *Aprtium* работает как открытая платформа для машинного перевода. В числе поддерживаемых — языков языки России: татарский, крымскотатарский и русский.

Автоматическая обработка естественного языка неизбежно порождает множественные варианты интерпретации одной и той же лексической единицы, что связано с феноменом грамматической омонимии. Разрешение такой неоднозначности представляет собой одну из ключевых задач компьютерной лингвистики. Ввиду колоссального объема современных корпусов ручное устранение омонимии практически неосуществимо, что стимулирует разработку автоматизированных методов обработки. В этом направлении ведутся интенсивные исследования, ориентированные на создание специализированных программных комплексов, использующих предопределенные правила, статистические алгоритмы или предобученные нейронные модели для повышения точности разметки (Сайхунов, Хусаинов, Ибрагимов, 2017: 382).

Морфологический анализатор *Apertium*, интегрированный в ПКТЯ, обладает встроенным механизмом разрешения определенных типов грамматической неоднозначности на основе системы формализованных правил. Дополнительным преимуществом данного программного решения является использование унифицированного набора тегов, применимого к широкому спектру языков, особенно относящихся к одной языковой группе. Это способствует созданию корпусов со схожей структурой морфологической аннотации, что значительно упрощает проведение межъязыковых сопоставительных исследований и сравнительного анализа лексико-грамматических закономерностей.

[29] Apertium. URL: https://www.apertium.org/index.rus.html#?dir=rus-bel&q= (дата обращения: 17.08.2025).

Дополнительные сервисы

Как правило, при разработке определенного корпуса преследуется решение конкретных задач фундаментального или практического плана. Однако, наблюдения, сделанные в процессе изучения ПКТЯ, указывают на то, что структурированные корпусные данные могут быть адаптированы для решения весьма широкого спектра проблем. Наличие масштабных аннотированных массивов текстовой информации позволяет авторам ПКТЯ интегрировать его ресурсы в различные технологические разработки и лингвистические приложения. Ниже представлены некоторые направления практического приложения рассматриваемого корпусного инструмента.

1. Онлайн-система проверки правописания, основанная на корпусных данных, позволяющая автоматизировать процесс орфографического контроля и корректировки ошибок при обучении татарскому языку или при иной проверке текстов на татарском языке.

2. Автоматическая генерация разнообразных статистических данных, включая частотные распределения букв, биграмм, слов, лемм и n-грамм, что предоставляет ценные сведения для количественного анализа закономерностей в татарском языке.

3. Создание тезауруса на основе дистрибутивной семантики, включающего векторные представления лексических единиц, сформированные с использованием технологий машинного обучения, таких как word2vec. В данном разделе разработаны 3 типа поиска: а) семантически схожие лексемы; б) аналогия лексем; в) сходства двух лексем по тем или иным признакам. Пример аналогии приведен на рис. 4.

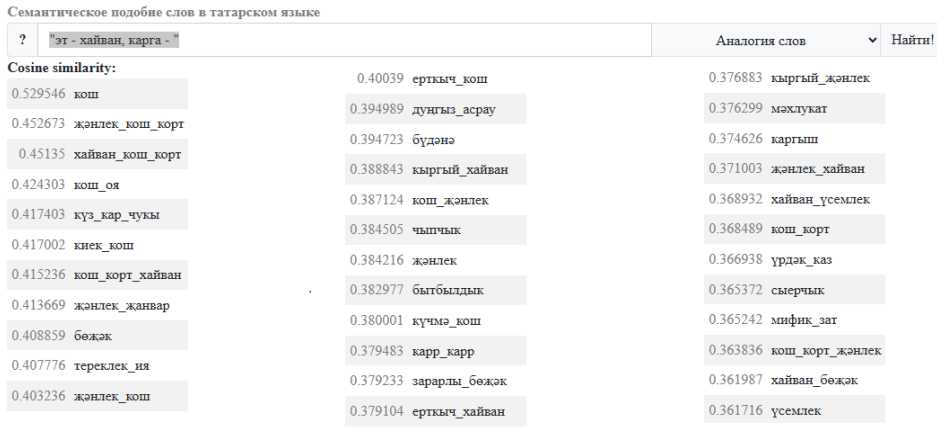

Рис. 4. Тезаурус. Пример типа поиска «аналогия слов»
Источник: Письменный корпус татарского языка.
URL: https://www.corpus.tatar/index.php?of=search/names.php (дата обращения: 12.08.2025).

Fig. 4. Thesaurus. Example of the search type **'**Word analogy**'**
Source: Corpus of written Tatar.
Retrieved 12 August 2025, from: https://www.corpus.tatar/index.php?of=search/names.php

4. Разработка системы синтеза татарской речи, предназначенной для создания различных тифлоинструментов и реализованной в Республиканской специальной библиотеке для слепых и слабовидящих[30] (рис. 5).

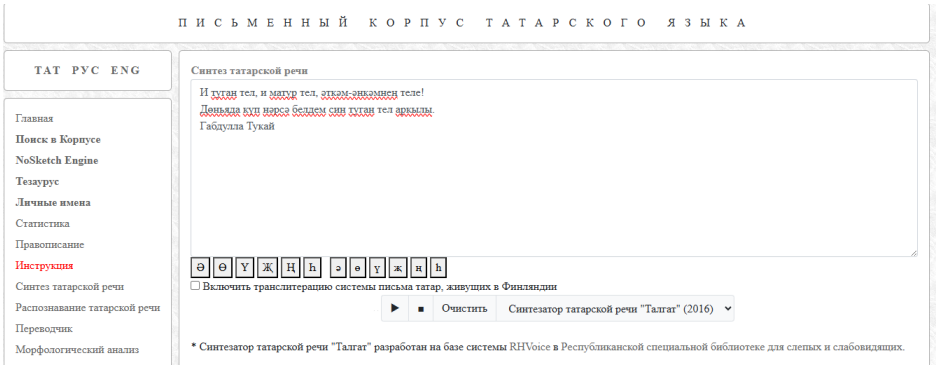

Рис. 5. Синтез татарской речи

Источник: Письменный корпус татарского языка. URL: https://www.corpus.tatar/index.php?of=search/names.php (дата обращения: 12.08.2025).

Fig. 5. Tatar Text-To-Speech

Source: Corpus of written Tatar. Retrieved 12 August 2025, from: https://www.corpus.tatar/index.php?of=search/names.php

5. Интеграция корпуса в систему распознавания татарской речи *Common Voice*[31], способствующей совершенствованию технологий автоматической транскрипции и взаимодействия с голосовыми интерфейсами (рис. 6).

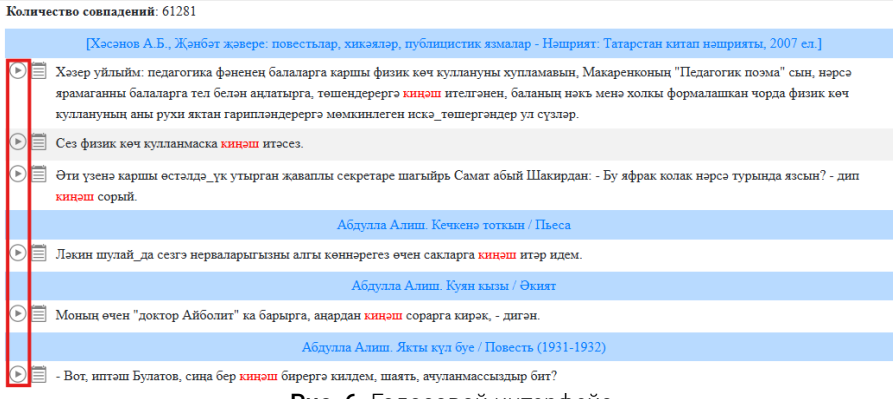

Рис. 6. Голосовой интерфейс

Источник: Письменный корпус татарского языка. URL: https://www.corpus.tatar/index.php?of=search/names.php (дата обращения: 12.08.2025).

Fig. 6. Voice interface

Source: Corpus of written Tatar. Retrieved 12 August 2025, from: https://www.corpus.tatar/index.php?of=search/names.php

[30] ГБУК Республики Татарстан «Республиканская специальная библиотека для слепых и слабовидящих имени Ш.К. Еникеева». URL: https://rsbsrt.ru/ (дата обращения: 18.07.2025).
[31] Common Voice. URL: https://commonvoice.mozilla.org/tt (дата обращения: 19.07.2025).

6. Оценка покрываемости морфологического анализатора *Apertium*, включающая выявление нерегулярных словоформ, автоматическое пополнение словарных ресурсов, а также оптимизацию системы правил для обучения технологий анализа естественного языка.

Некоторые из перечисленных проектов реализованы в формате веб-сервисов, что способствует расширению пользовательской аудитории и повышает прикладную значимость корпусных исследований, способствуя развитию цифровых лингвистических технологий.

Вклад Письменного корпуса татарского языка в фундаментальные и прикладные исследования

Письменный корпус татарского языка представляет собой не только технически значимый цифровой ресурс, но и платформу, имеющую фундаментальную ценность, которая активно используется в широком спектре лингвистических и междисциплинарных исследований.

На материале корпуса выполняются исследования в следующих направлениях:

• морфологический анализ и автоматическая обработка текста, включая разработку алгоритмов морфологического разборщика и анализ продуктивных словообразовательных моделей (Сайхунов, Хусаинов, Ибрагимов, 2018);

• изучение лексики и грамматики татарского языка, включая частотный анализ, синтаксические конструкции, словообразовательные модели, синонимию и фонологию (Ибраһимов, Сәйхунов, 2016; Ибрагимова, 2020; Кузнецов, 2023; Galieva, Vavilova, Gafarova, 2017; Luutonen, Moisio, Daher, 2017);

• исследование татарского фольклора и литературы, охватывающее жанровую эволюцию, особенности художественной речи и национально-культурные коды (Сайхунов, Ибрагимов, Галиуллин, 2018);

• социолингвистические исследования, посвященные анализу языковой ситуации, сохранению и трансформации татарской идентичности, а также вопросам языковой политики в условиях многоязычия (Ибрагимов, Сайхунов, 2015);

• цифровые и прикладные проекты, включая участие в создании систем распознавания и синтеза татарской речи на базе корпуса (Сайхунов, 2010);

• методология корпусной лингвистики, охватывающая вопросы создания, структурирования и использования национальных корпусов, в т.ч. на ЯНР (Ибрагимов, Сайхунов, Салимзянов, 2012; Гатиатуллин, Мухамедшин, Прокопьев, Сулейманов, 2024: 543).

Научная и практическая значимость названных исследований заключается в том, что они обеспечивают: 1) формирование репрезентативной эмпирической

базы для описания и нормализации татарского языка; 2) развитие технологий автоматической обработки тюркских языков, ранее недостаточно представленных в цифровом пространстве; 3) поддержку процессов цифровизации языков народов России; 4) интеграцию татарского языка в международные исследовательские проекты в области компьютерной лингвистики.

Перспективы развития и использования Письменного корпуса татарского языка

Развитие ПКТЯ имеет стратегическое значение как для лингвистической науки, так и для государственной языковой политики, направленной на сохранение и цифровую поддержку языков народов России. С учетом вызовов времени и международных инициатив в сфере языкового многообразия дальнейшая работа над корпусом открывает широкие перспективы.

Основные направления развития корпуса могут включать:

• расширение текстовой базы, включая добавление произведений современной художественной литературы, научных публикаций в разных областях знаний, публицистики и официальных документов, что повысит тематическую и стилевую репрезентативность корпуса;

• развитие лингвистической разметки, в частности — совершенствование морфологического и синтаксического анализа, а также внедрение элементов семантической и прагматической аннотации;

• создание специализированных подкорпусов, отражающих территориальные диалекты, возрастные и гендерные особенности, профессиональные и жанровые различия употребления языка;

• разработка электронных лингвистических ресурсов, таких как частотные словари, тезаурусы и обучающие базы, предназначенные для преподавания татарского языка и автоматической обработки текстов (Ибрагимов, Сайхунов, 2018: 73).

Перспективы применения корпуса в будущем видятся в следующих плоскостях:

• цифровизация образования, включая интеграцию корпуса в онлайн-платформы для обучения татарскому языку;

• машинный перевод и распознавание речи, где корпус будет служить основой для обучения моделей искусственного интеллекта;

• лексикографические и грамматические исследования, способствующие созданию новых нормативных справочников татарского языка;

• поддержка культурной политики, включая анализ языковой динамики и адаптацию языка к условиям цифровой среды;

• проектная деятельность в области Digital Humanities, ориентированная на разработку мультимодальных и межъязыковых ресурсов.

Деятельность по развитию ПКТЯ полностью соответствует задачам, поставленным в рамках Международного десятилетия языков коренных народов (2022–2032)[32], инициированного Генеральной Ассамблей ООН и координируемого ЮНЕСКО. Инициатива направлена на сохранение, развитие и распространение языков, находящихся, в т.ч. под угрозой исчезновения.

Заключение

ПКТЯ представляет собой масштабный лингвистический ресурс, структурированный в соответствии с современными требованиями, стандартами и традициями корпусной лингвистики. Он включает тексты различных жанров, стилей и исторических периодов, охватывая широкий спектр письменной речевой практики. Корпус аннотирован с применением морфологического анализатора *Apertium*, что обеспечивает высокую точность разметки и возможность формального анализа грамматических структур.

Организационная система корпуса обеспечивается метаразметкой по ключевым экстралингвистическим параметрам: авторство, дата, жанровая и стилистическая принадлежность, источник и формат публикации. Несмотря на использование собственного формата представления данных, структура корпуса позволяет эффективно решать широкий круг исследовательских задач в области лексикографии, грамматического анализа, исторической лингвистики, межъязыкового сопоставления и обработки естественного языка.

Благодаря открытой архитектуре и машиночитаемому формату ПКТЯ интегрирован в ряд прикладных сервисов: автоматическую орфографическую проверку, генерацию статистических параметров, построение тезаурусов на основе word2vec, синтез и распознавание татарской речи. Эти разработки способствуют цифровизации татарского языка и его активному использованию в новых технологических контекстах.

Корпус продолжает развиваться. Планируется расширение жанрового охвата, улучшение морфологической аннотации, переход к стандартизированным форматам хранения и создание подкорпусов для региональных вариантов языка. Вектор развития согласуется с международной повесткой, связанной с сохранением языкового многообразия, в т.ч. в рамках Международного десятилетия языков коренных народов (2022–2032). Письменный корпус татарского языка представляет собой не только научный, но и культурно-стратегический проект, направленный на поддержку и продвижение национального языка в цифровую эпоху.

[32] Международное десятилетие языков коренных народов (2022–2032) // UNESCO. URL: https://www.unesco.org/ru/decades/indigenous-languages (дата обращения: 15.08.2025).

Список литературы

Алюнина Ю.М. Цифровые технологии в переводе. СПб. : Лань, 2025. 144 с.

Белоруссова С.Ю. Киберэтнография: методология и технология // Этнография. 2021. № 3 (13). С. 123–145. https://doi.org/10.31250/2618-8600-2021-3(13)-123-145 EDN: DXBUUJ

Богородицкий В.А. Введение в татарское языкознание в связи с другими тюркскими языками / под ред. Н.К. Дмитриева. Казань : Татгосиздат, 1953. 220 с.

Гатиатуллин А.Р., Мухамедшин Д.Р., Прокопьев Н.А., Сулейманов Д.Ш. Электронный корпус татарского языка на базе модели лингвистических графов знаний // Онтология проектирования. 2024. Т. 14. № 4 (54). С. 542–554. https://doi.org/10.18287/2223-9537-2024-14-4-542-554 EDN: FXVFET

Ибрагимов Т.И., Сайхунов М.Р. Письменный корпус татарского языка: структурные и функциональные характеристики // Актуальные проблемы диалектологии языков народов России : материалы XIV Всеросс. науч. конф., Уфа, 20–22 ноября 2014 г. / отв. ред. Ф.Г. Хисамитдинова. Уфа : Институт истории, языка и литературы Уфимского научного центра РАН, 2014. С. 261–264. EDN: TIEBGB

Ибрагимов Т.И., Сайхунов М.Р. Языковое состояние этнической общности на материале Письменного корпуса татарского языка // Tatarica. 2015. № 1 (4). С. 22–27. EDN: UNYXIL

Ибраһимов Т.И., Сәйхунов М.Р. Хәзерге татар сөйләм теле: сузык авазлар составы // Фәнни Татарстан. 2016. № 3. С. 35–47. EDN: XRPKMH

Ибрагимова Э.Р. К вопросу о референтном и атрибутивном употреблении наименований лица в предложениях тождества в английском и татарском языках // Филология и культура. 2020. № 4 (62). С. 36–42. https://doi.org/10.26907/2074-0239-2020-62-4-36-42 EDN: RXAZHS

Кузнецов М.Ю. Алгоритм нахождения основы татарского глагола по его инфинитиву (орфографическо-грамматические аспекты) // Многоязычие в образовательном пространстве. 2023. Т. 15. № 2 (17). С. 192–202. https://doi.org/10.35634/2500-0748-2023-15-2-192-202 EDN: YUGYOJ

Родионова А.П., Пеллинен Н.А. Корпусная лингвистика и марафон записей вепсской и карельской речи как инструмент популяризации прибалтийско-финских языков Карелии // Macrosociolinguistics and Minority Languages. 2024. Т. 2. № 2. С. 115–130. https://doi.org/10.22363/2312-797X2024-2-2-115-130 EDN: IPVMVI

Сайхунов М.Р. Ритмо-темпоральные характеристики татарского языка в плане автоматического синтеза речи : автореф. дис. … канд. филол. наук. Казань, 2010. 26 с. EDN: QGVUYR

Сайхунов М.Р., Ибрагимов Т.И., Галиуллин К.Р. Корпус татарской художественной литературы // Традиционная культура народов Поволжья : материалы IV Всеросс. науч.-практ. конф. с междунар. участием. Казань : ИХЛАС, 2018. С. 370–377. EDN: USSRQS

Сайхунов М.Р., Хусаинов Р.Р., Ибрагимов Т.И. Система сложного морфологического поиска в Письменном корпусе татарского языка // Традиционная культура тюркских народов в изменяющемся мире : материалы I Междунар. науч. конф. Казань : Ак Буре, 2017. С. 382–385. EDN: YSRGAQ

Сайхунов М.Р., Хусаинов Р.Р., Ибрагимов Т.И. Сложности при создании текстового корпуса объемом более 400 млн токенов // Финно-угорский мир в полиэтничном пространстве России: культурное наследие и новые вызовы : сб. статей по материалам VI Всеросс. науч. конф. финно-угроведов, Ижевск, 04–07 июня 2019 г. Ижевск : Изд-во Анны Зелениной, 2019. С. 548–554. EDN: LSYZBT

Сайхунов М.Р., Хусаинов Р.Р., Ибрагимов Т.И. Эволюция систем поиска в Письменном корпусе татарского языка // Языковые контакты народов Поволжья и Урала : сб. статей XI Междунар. симпозиума, Чебоксары, 21–24 мая 2018 г. Чебоксары : Чувашский государственный университет им. И.Н. Ульянова, 2018. С. 314–319. EDN: XVTCWD

Galieva A., Vavilova Z., Gafarova V. Developing Tatar corpus-based dictionaries for educational purposes // INTED2017 Proceedings. Valencia : INTED, 2017. P. 9014–9022. https://doi.org/10.21125/inted.2017.2131

Luutonen J., Moisio A., Daher O. Finnish Tatars and the trilingual Tatar-Finnish dictionary // Turcic Languages. 2017. Vol. 21. № 2. P. 266–280. https://doi.org/10.13173/TL/2017/2/266

References

Alyunina, Yu.M. (2025). *Tsifrovye tekhnologii v perevode [Digital technologies in translation]*. Lan' Publ. (In Russ.).

Belorussova, S. Yu. (2021). Cyberethnography: Methodology and technology. *Etnografia*, (3), 123–145. (In Russ.). https://doi.org/10.31250/2618-8600-2021-3(13)-123-145 EDN: DXBUUJ

Bogoroditsky, V.A., Dmitriev, N., ed. (1953). *Vvedenie v tatarskoe yazykoznanie v svyazi s drugimi tyurkskimi yazykami [Introduction to Tatar Linguistics in Relation to Other Turkic Languages].* Kazan: Tatgosizdat publ. (In Russ.).

Galieva, A., Vavilova, Z., & Gafarova, V. (2017). Developing Tatar corpus-based dictionaries for educational purposes. *INTED2017 Proceedings*, 9014–9022. https://doi.org/10.21125/inted.2017.2131

Gatiatullin, A.R., Mukhamedshin, D.R., Prokopyev, N.A., & Suleymanov, D. Sh. (2024). Electronic corpus of the Tatar language based on the model of linguistic knowledge graphs. *Ontology of Designing*, *14*(4), 542–554. (In Russ.). https://doi.org/10.18287/2223-9537-2024-14-4-542-554 EDN: FXVFET

Ibragimov, T.I., & Saikhunov, M.R. (2014). The written corpus of the Tatar language: Structural and functional characteristics. In F.G. Khisamitdinova (ed.). *Proceedings of the 14th All-Russian scientific conference «Current issues in the dialectology of the languages of the peoples of Russia»,* 261–264. Ufa: Federal State Budgetary Institution of Science, Institute of History, Language and Literature, Ufa Scientific Center of the Russian Academy of Sciences. (In Russ.). EDN: TIEBGB

Ibragimov, T.I., & Saikhunov, M.R. (2015). The language status of an ethnic community (on the material of the Tatar language written corpus). *Tatarica*, (1), 22–27. (In Russ.). EDN: UNYXIL

Ibrahimov, T.I., & Saykhunov, M.R. (2016). The modern Tatar spoken language: The structure of vowel sounds. *Fənni Tatarstan*, (3), 35–47. (In Tatar). EDN: XRPKMH

Ibragimova, E.R. (2020). On the reference and attributive use of personal names in English and Tatar identical utterances. *Philology and Culture,* (4), 36–42. (In Russ.). https://doi.org/10.26907/2074-0239-2020-62-4-36-42 EDN: RXAZHS

Kuznetsov, M. Yu. (2023). Algorithm for finding the base of the Tatar verb by its infinitive (orthographic-grammatical aspects). *Russian Journal of Multilingualism and Education, 15*(2), 192–202. (In Russ.). https://doi.org/10.35634/2500-0748-2023-15-2-192-202 EDN: YUGYOJ

Luutonen, J., Moisio, A., & Daher, O. (2017). Finnish Tatars and the trilingual Tatar-Finnish dictionary. *Turcic Languages*, *21*(2), 266–280. https://doi.org/10.13173/TL/2017/2/266

Rodionova, A.P., & Pellinen, N.A. (2024). The corpus linguistics and the marathon of recordings of Vepsian and Karelian speech as a tool for popularizing the Baltic-Finnish languages of Karelia. *Macrosociolinguistics and Minority Languages, 2*(2), 115–130. (In Russ.). https://doi.org/10.22363/2312-797X2024-2-2-115-130 EDN: IPVMVI

Saikhunov, M.R. (2010). *Ritmo-temporal'nye kharakteristiki tatarskogo yazyka v plane avtomaticheskogo sinteza rechi [Rhythmic and temporal characteristics of the Tatar language in the context of automatic speech synthesis]* [Dissertation abstract]. Kazan: Kazan State University. (In Russ.). EDN: QGVUYR

Saikhunov, M.R., Ibragimov, T.I., & Galiullin, K.R. (2018). The Corpus of Tatar fiction literature. *Traditional culture of the peoples of the Volga Region: Proceedings of the 4th All-Russian scientific and practical conference with international participation*, p. 370–377. Kazan: Ikhlas publ. (In Russ.). EDN: USSRQS

Saikhunov, M.R., Khusainov, R.R., & Ibragimov, T.I. (2017). The system of advanced morphological search in the written corpus of the Tatar language. *Traditional culture of Turkic peoples in a changing world: Proceedings of the 1st international scientific conference*, p. 382–385. Kazan: Ak Bure publ. (In Russ.). EDN: YSRGAQ

Saikhunov, M.R., Khusainov, R.R., & Ibragimov, T.I. (2018). Search systems evolution in the corpus of written Tatar language. *Language contacts of the peoples of the Volga Region and the Urals: Proceedings of the 11th international symposium*, p. 314–319. Chuvash State University named after N. Ulyanov. (In Russ.). EDN: XVTCWD

Saikhunov, M.R., Khusainov, R.R., & Ibragimov, T.I. (2019). Challenges in creating a text corpus exceeding 400 million tokens. *The Finno-Ugric World in the Multiethnic Space of Russia: Cultural Heritage and New Challenges: Proceedings of the 6th All-Russian conference on Finno-Ugric studies*, p. 548–554. Izhevsk, Anna Zelenina publ. (In Russ.). EDN: LSYZBT

Сведения об авторе:
КЕНЕШБЕКОВА Шахдар Нурдиновна, аспирант кафедры социальной педагогики Института иностранных языков, Российский университет дружбы народов, Российская Федерация, 117198, г. Москва, ул. Миклухо-Маклая, д. 6; старший преподаватель кафедры лингвистики и межкультурной коммуникации Гуманитарного института, Российский новый университет, Российская Федерация, 105005, Москва, ул. Радио, д. 22. *Научные интересы*: корпусная лингвистика, компьютерная лингвистика, тюркские языки, сопоставительные исследования, методика преподавания иностранных языков.
E-mail: keneshbekova.sh@yandex.ru
ORCID: 0009-0002-3599-2065
SPIN-код: 8204-2172

Bionote:
Shakhdar N. KENESHBEKOVA, a PhD student at the Department of Social Pedagogy, Institute of Foreign Languages, RUDN University, 6 Miklukho-Maklaya St., Moscow, 117198, Russian Federation; Senior Lecturer at the Department of Linguistics and Intercultural Communication, Institute of Humanities, Russian New University, 22 Radio St., Moscow, 105005, Russian Federation. *Research interests:* corpus linguistics, computational linguistics, Turkic languages, contrastive studies, methods of teaching foreign languages.
E-mail: keneshbekova.sh@yandex.ru
ORCID: 0009-0002-3599-2065
SPIN-code: 8204-2172