УДК 811.512.111′322
ББК 81.635.1+32.972

Научная статья

КОНЦЕПЦИЯ ЭЛЕКТРОННОГО КОРПУСА ЧУВАШСКОГО ЯЗЫКА: ТЕХНИЧЕСКИЙ АСПЕКТ

Д. М. Леонтьев
Чувашский государственный институт гуманитарных наук,
г. Чебоксары, Россия
denis.lentev@mail.ru

Аннотация. В статье рассматривается процесс создания национального корпуса чувашского языка, при этом внимание автора акцентировано на его технической стороне. В качестве технологической платформы был выбран фреймворк Laravel, обладающий рядом преимуществ, включая гибкость, отсутствие лицензионных ограничений, удобство работы с большими объемами данных, наличие веб-интерфейса и интеграцию с аналитическими инструментами. Для обеспечения унификации данных был введен стандарт Unicode, позволяющий корректно обрабатывать специфические чувашские символы. На этапе формирования корпуса был осуществлен сбор текстов из открытых источников, создана база данных и проведена обработка текстов с помощью макросов Microsoft Word и системы фильтров для приведения символов к единому стандарту. Разработаны алгоритмы выявления и замены нестандартных символов. Разработанный электронный корпус чувашского языка не только соответствует современным технологическим требованиям, но и открывает широкие возможности для лингвистических исследований и прикладных задач в сфере гуманитарных наук.

Ключевые слова: электронный корпус, чувашский язык, фреймворк, Laravel, база данных, Unicode, фильтрация текста, поисковая система

Для цитирования: Леонтьев Д. М. Концепция электронного корпуса чувашского языка: технический аспект // Современная гуманитаристика. 2025. Т. 1. № 2. С. 75 — 90

Scientific article

THE CHUVASH LANGUAGE ELECTRONIC CORPUS CONCEPT: TECHNICAL ASPECT

D. M. Leontiev
Chuvash State Institute of Humanities,
г. Cheboksary, Russia
denis.lentev@mail.ru

Abstract. The article deals with the process of national corpus creation of the Chuvash language, particularly its technical side. Laravel framework has been chosen as a technological platform having a number of advantages it includes: flexibility, absence of license restrictions, convenience of working with large amounts of data, availability of web-interface and integration with analytical tools. To ensure data unification, the Unicode standard was introduced allowing to correctly processing specific Chuvash symbols. At the corpus building stage, texts have been collected from open sources, a database has been created and texts using macros in Microsoft Word and a system of filters to bring the symbols to a single standard have been processed. Algorithms for identifying and replacing non-standard symbols have been developed. The developed electronic corpus of the Chuvash language both meets modern technological requirements and opens wide opportunities for linguistic research and applied tasks in the field of humanities.

Keywords: electronic corpus, the Chuvash language, framework, Laravel, database, Unicode, text filtering, search engine

For citation: Leontiev D. M. The Chuvash language electronic corpus concept: technical aspect // Modern humanities. 2025;1(2):75 — 90

Ăслăлăх статйи

ЧĂВАШ ЧĔЛХИН ЭЛЕКТРОНЛĂ КОРПУСĔН КОНЦЕПЦИЙĔ: ТЕХНИКА АСПЕКЧĔ

Д. М. Леонтьев
Чăваш патшалăх гуманитари ăслăлăхĕсен институчĕ,
Шупашкар, Раççей
denis.lentev@mail.ru

Аннотаци. Статьяра чăваш чĕлхин наци корпусне тăвас процеса пăхса тухнă, çав хушăрах ытларах унăн техника ыйтăвĕсене тимленĕ. Технологи платформисенчен пахалăх енчен ыттисенчен лайăхраххине — Laravel фреймворка суйласа илнĕ. Лайăх енсен шутĕнче: куçăмлăх, лицензи чаракĕсем çукки, пысăк калăпăшлă даннăйсемпе ĕçлеме меллĕ пулни, веб-интерфейс тата аналитика хатĕрĕсемпе интеграцилени пурри. Даннăйсене унификацилеме чăваш символĕсене тӳрремĕн кĕртме май паракан Unicode стандартпа усă курнă. Корпуса йĕркеленĕ май уçă çăл куçсенчи текстсене пухнă, даннăйсен пуххине йĕркеленĕ Microsoft Word макросĕсемпе символĕсене тата фильтрсен тытăмĕпе усă курса текстсене пĕр стандартпа туса вырнаçтарнă. Стандартлă мар символсене палăртмалли тата улăштармалли алгоритмсем хатĕрленĕ. Институтра хатĕрленĕ чăваш чĕлхин электронлă корпусĕ хальхи технологи требованийĕсемпе килĕшсе тăрать, унсăр пуçне лингвистика тĕпчевĕсем тума тата гуманитари ăслăлăхĕсен сферинче прикладной задачăсене татса пама анлă майсем уçать.

Тĕп сăмахсем: электронлă корпус, чăваш чĕлхи, фреймворк, Laravel, даннăйсен бази, Unicode, текст фильтрацийĕ, шырав системи

Цитатăлама: Леонтьев Д. М. Чăваш чĕлхин электронлă корпусĕн концепцийĕ: техника аспекчĕ // Хальхи гуманитаристика. 2025. 1 т. 2 №. С. 75 — 90

Введение

В условиях активной цифровизации гуманитарных наук создание национальных корпусов языков становится важной задачей для лингвистов, программистов и исследователей. Электронный корпус представляет собой структурированную базу текстов, позволяющую изучать язык на разных уровнях — от морфологии и синтаксиса до социолингвистических аспектов. Разработка национального корпуса чувашского языка играет ключевую роль в сохранении и изучении языкового наследия, а также в создании инструментов для автоматической обработки текстов.

Материалы и методы

Для создания национального корпуса чувашского языка были использованы цифровые тексты различных жанров, прошедшие техническую обработку. Работа включала сбор, форматирование и очистку данных, разработку структуры базы MySQL, а также создание веб-приложения на Laravel и Bootstrap 5. Применялись аналитический, фактологический, описательный и сравнительный методы, обеспечивающие точность, структурирование и сопоставление подходов. Разработка велась в Visual Studio Code с использованием MySQL, OpenServer, PHP, JavaScript, Git и Composer. В результате проведенных работ создана платформа для изучения чувашского языка и автоматизированного анализа текстов.

Результаты исследования и их обсуждение

В современных условиях цифровизации гуманитарного знания идет активная работа по созданию лингвистических корпусов, представляющих собой струк-

турированную коллекцию текстов одного или нескольких языков, хранящихся в электронной базе данных (БД) и предназначенных для научных исследований.

Первые электронные корпусы были разработаны в 60 — 70-е гг. XX в. преимущественно на материале английского языка (стандартный корпус американского английского языка Университета Брауна и Корпус британского английского языка Ланкастер-Осло-Берген), но вскоре появились корпусы и на базе других языков [1, с. 8]. В настоящее время созданы и функционируют электронные корпусы для многих национальных языков. Первая общедоступная версия Национального корпуса русского языка появилась в интернете в 2004 г., в дальнейшем стали разрабатываться электронные корпусы и для других языков народов нашей страны. Например, успешно развиваются национальные корпусы татарского, башкирского, марийского, калмыцкого, бурятского и других языков.

В данной статье мы рассматриваем технический аспект концепции электронного корпуса чувашского языка. Создание национального корпуса необходимо для решения задач как теоретической, так и прикладной лингвистики. Корпус предоставляет возможность оперативно находить слова с нужными исследователю признаками. Такие поисковые задачи невозможно осуществить ни с помощью обычного текстового редактора, ни с помощью интернета. Корпус позволяет выявлять частотность слов и выражений, изучать грамматические структуры, подтверждать лингвистические гипотезы, анализировать диалектные и социально-лексические вариации, проводить исследования языковых изменений и т. д. Он также может быть использован в исследованиях в области социолингвистики, психолингвистики, переводоведения, литературоведения и других научных областях.

Различные аспекты создания электронного корпуса чувашского языка исследовались П. В. Желтовым [3; 4], В. П. Желтовым [5], А. Р. Губановым [2]. Большой вклад в развитие корпуса чувашского языка внесли Н. А. Плотников [6] и А. Н. Антонов [7].

Реализация технического аспекта концепции Национального корпуса чувашского языка включает в себя несколько этапов. Чувашский государственный институт гуманитарных наук приступил к технической разработке корпуса в октябре 2023 г. На первоначальном этапе одной из ключевых задач явился выбор подходящей технологической платформы для корпуса, на которой будет осуществляться обработка, хранение и анализ текстовых данных. Были изучены различные готовые специализированные программы для создания электронных корпусов языков — Sketch Engine [12], AntConc [10], Voyant Tools [13], CQPweb [8], NoSketch Engine [11] и др. Выбор был сделан в пользу фреймворка Laravel, занимающего ведущие позиции в мировом рейтинге PHP-фреймворков [9]. Основные его преимущества по сравнению с другими программами заключаются в следующем:

А) Гибкость и настройка под индивидуальные требования. Большинство готовых программ для разработки электронных корпусов (например, Sketch Engine, AntConc, Voyant Tools) включают в себя лишь фиксированный набор функций (например, анализ частотности и визуализацию данных), что подходит для базовых задач, но ограничивает возможность каких-либо дальнейших расширений. Если нужно добавить специфическую функциональность (к примеру, интеграцию с уникальными языковыми модулями или нестандартные алгоритмы анализа), то это либо совсем невозможно, либо требует значительных усилий. В отличие от этих программ, платформа Laravel содержит не только набор универсальных ин-

струментов, но она способна еще полностью адаптировать функционал под потребности конкретного проекта. Например, в электронном корпусе языка можно разработать уникальные инструменты для морфологического анализа, добавить функции автоматического исправления текста или сложные механизмы поиска. Гибкая архитектура платформы Laravel позволяет легко масштабировать проект — включать новые модули, изменять существующие или интегрировать сторонние сервисы (машинное обучение, искусственный интеллект, платежные системы, системы аналитики и мониторинга, социальные сети и т. д.);

Б) Отсутствие зависимости от сторонних ограничений. Готовые программы для создания электронных корпусов (Sketch Engine, CQPweb или NoSketch Engine) имеют высокую стоимость, лицензионные ограничения, они зависят от коммерческих обновлений и требуют значительных усилий для настройки и интеграции. Laravel полностью открытый фреймворк, который не накладывает ограничений на его использование, есть возможность контролировать каждую часть его системы — от БД до пользовательского интерфейса. Продукт, созданный на платформе Laravel, является полностью автономным и бесплатным, независимым от обновлений сторонних разработчиков;

В) Удобство работы с большими объемами данных. Одна часть готовых программ для разработки электронных корпусов (AntConc или Voyant Tools) рассчитана на локальный анализ текстов, их производительность ограничена при работе с большими корпусами, другая (Sketch Engine или CQPweb) лучше справляется с большими объемами данных, однако требует предварительной настройки серверов и конфигурации. Платформа Laravel позволяет легко работать с большими объемами данных. Благодаря возможности быстрого подключения современных БД (MySQL, PostgreSQL, MongoDB и др.) на этой платформе можно реализовать оптимизированные алгоритмы обработки, которые эффективно работают даже с десятками миллионов записей;

Г) Наличие веб-доступа и интерфейса. Зачастую готовые программы для создания электронных корпусов либо совсем не имеют веб-интерфейса (например, AntConc), либо требуют сложной его установки (CQPweb и др.). Их интерфейсы ориентированы в основном на специалистов, у неподготовленных пользователей могут вызвать затруднения. Фреймворк Laravel предоставляет возможность создания удобного веб-интерфейса для электронного корпуса, доступного как ученым-лингвистам, так и неспециалистам. Любой интернет-пользователь может открыть сайт корпуса в браузере, ввести слово или фразу и сразу получить нужные данные, не разбираясь в сложных настройках;

Д) Возможность интеграции с аналитическими инструментами. Готовые программы для создания электронных корпусов, например, Sketch Engine или Voyant Tools, предоставляют встроенные функции анализа, их возможности ограничены заранее определенным набором инструментов. В отличие от них, платформа Laravel позволяет интегрировать сторонние аналитические системы, включая библиотеки для обработки текста (например, NLTK, SpaCy) и решения для машинного обучения (TensorFlow, PyTorch и др.). Это позволяет добавить функции, которые недоступны в готовых программах, такие как автоматическое обучение на основе корпуса данных или динамическую адаптацию алгоритмов под новые данные;

Е) Наличие долгосрочной поддержки и масштабируемости. Электронный корпус языка, созданный на основе готовой программы, зависим от сторон-

них разработчиков, которые имеют возможность ограничить долгосрочную его поддержку. Если программное обеспечение корпуса устаревает или разработка прекращается, то его использование становится весьма проблематичным. Во фреймворке Laravel есть возможность полностью самостоятельно управлять разработанным продуктом. При необходимости его можно обновлять, масштабировать и адаптировать к новым требованиям. Имеется большая поддержка сообщества разработчиков Laravel с точки зрения регулярных обновлений, исправления ошибок, а также создания новых функций и инструментов, которые могут быть использованы в дальнейшей разработке электронного корпуса.

Вместе с тем необходимо отметить, что платформа Laravel имеет и ряд недостатков. Так, при работе с большими массивами текста требуются мощные серверные ресурсы, при увеличении нагрузки на систему также может возникнуть необходимость в переходе на специализированные решения для хранения данных, такие как NoSQL-БД. Кроме того, фреймворк Laravel имеет сложную архитектуру, что затрудняет его освоение для программистов с небольшим опытом работы.

На втором этапе разработки национального корпуса чувашского языка для унификации всех данных был введен единый юникод-стандарт символов (Unicode). Он представлен следующим массивом S1:

S1 = ['а', # (U+0430) 'б', # (U+0431) 'в', # (U+0432) 'г', # (U+0433) 'д', # (U+0434) 'е', # (U+0435) 'ж', # (U+0436) 'з', # (U+0437) 'и', # (U+0438) 'й', # (U+0439) 'к', # (U+043A) 'л', # (U+043B) 'м', # (U+043C) 'н', # (U+043D) 'о', # (U+043E) 'п', # (U+043F) 'р', # (U+0440) 'с', # (U+0441) 'т', # (U+0442) 'у', # (U+0443) 'ф', # (U+0444) 'х', # (U+0445) 'ц', # (U+0446) 'ч', # (U+0447) 'ш', # (U+0448) 'щ', # (U+0449) 'ъ', # (U+044A) 'ы', # (U+044B) 'ь', # (U+044C) 'э', # (U+044D) 'ю', # (U+044E) 'я', # (U+044F) 'Ё', # (U+0401) 'ё', # (U+0451) 'А', # (U+0410) 'Б', # (U+0411) 'В', # (U+0412) 'Г', # (U+0413) 'Д', # (U+0414) 'Е', # (U+0415) 'Ж', # (U+0416) 'З', # (U+0417) 'И', # (U+0418) 'Й', # (U+0419) 'К', # (U+041A) 'Л', # (U+041B) 'М', # (U+041C) 'Н', # (U+041D) 'О', # (U+041E) 'П', # (U+041F) 'Р', # (U+0420) 'С', # (U+0421) 'Т', # (U+0422) 'Ё', # (U+0401) 'У', # (U+0423) 'Ф', # (U+0424) 'Х', # (U+0425) 'Ц', # (U+0426) 'Ч', # (U+0427) 'Ш', # (U+0428) 'Щ', # (U+0429) 'Ъ', # (U+042A) 'Ы', # (U+042B) 'Ь', # (U+042C) 'Э', # (U+042D) 'Ю', # (U+042E) 'Я', # (U+042F) 'a', # (U+0061) 'b', # (U+0062) 'c', # (U+0063) 'd', # (U+0064) 'e', # (U+0065) 'f', # (U+0066) 'g', # (U+0067) 'h', # (U+0068) 'i', # (U+0069) 'j', # (U+006A) 'k', # (U+006B) 'l', # (U+006C) 'm', # (U+006D) 'n', # (U+006E) 'o', # (U+006F) 'p', # (U+0070) 'q', # (U+0071) 'r', # (U+0072) 's', # (U+0073) 't', # (U+0074) 'u', # (U+0075) 'v', # (U+0076) 'w', # (U+0077) 'x', # (U+0078) 'y', # (U+0079) 'z', # (U+007A) 'A', # (U+0041) 'B', # (U+0042) 'C', # (U+0043) 'D', # (U+0044) 'E', # (U+0045) 'F', # (U+0046) 'G', # (U+0047) 'H', # (U+0048) 'I', # (U+0049) 'J', # (U+004A) 'K', # (U+004B) 'L', # (U+004C) 'M', # (U+004D) 'N', # (U+004E) 'O', # (U+004F) 'P', # (U+0050) 'Q', # (U+0051) 'R', # (U+0052) 'S', # (U+0053) 'T', # (U+0054) 'U', # (U+0055) 'V', # (U+0056) 'W', # (U+0057) 'X', # (U+0058) 'Y', # (U+0059) 'Z', # (U+005A) '0', # (U+0030) '1', # (U+0031) '2', # (U+0032) '3', # (U+0033) '4', # (U+0034) '5', # (U+0035) '6', # (U+0036) '7', # (U+0037) '8', # (U+0038) '9', # (U+0039) '.', # (U+002E) ',', # (U+002C) ';', # (U+003B) ':', # (U+003A) '!', # (U+0021) '?', # (U+003F) '(', # (U+0028) ')', # (U+0029) '[', # (U+005B) ']', # (U+005D) '{', # (U+007B) '}', # (U+007D) '<', # (U+003C) '>', # (U+003E) '\', # (U+0027) '»', # (U+0022) '\\', # (U+005C) '/', # (U+002F) '|', # (U+007C) '-', # (U+002D) '+', # (U+002B) '=', # (U+003D) '*', # (U+002A) '_',

(U+005F) '~', # (U+007E) '&', # (U+0026) '@', # (U+0040) '#', # (U+0023) '$', # (U+0024) '%', # (U+0025) '^', # (U+005E) '«', # (U+00AB) '»', # (U+00BB) '–', # (U+2013) ' ', # (U+0020) '№', # (U+2116) '…', # (U+2026) '©', # (U+00A9) '•', # (U+00B7) 'Ç', # (U+04AA) 'ç', # (U+04AB) 'Ă', # (U+04D0) 'ă', # (U+04D1) 'Ĕ', # (U+04D6) 'ĕ', # (U+04D7) 'Ӳ', # (U+04F2) 'ӳ' # (U+04F3)].

Unicode необходим, во-первых, для приведения к единообразной форме текстов корпуса, взятых из разных источников и имеющих соответственно различные стили написания и кодировки (UTF-8, Windows-1251 и др.); во-вторых, для сохранения особенностей чувашского языка — уникальных специфических символов «Ă» (Unicode: U+04D0), «ă» (Unicode: U+04D1), «Ĕ» (Unicode: U+04D6), «ĕ» (Unicode: U+04D7), «Ç» (Unicode: U+04AA), «ç» (Unicode: U+04AB), «Ӳ» (Unicode: U+04F2), «ӳ» (Unicode: U+04F3); в-третьих, для правильного проведения морфологического анализа и синтаксического разбора (допустим, если в тексте используется нестандартная форма записи букв, то при морфологическом анализе программа может неправильно распознать форму слова); в-четвертых, для упрощения поисковых запросов; в-пятых, для обеспечения совместимости национального корпуса с любыми другими современными системами обработки текста.

На третьем этапе создания электронного корпуса чувашского языка был осуществлен сбор разнообразных текстов, пригодных для проведения лингвистических исследований, а затем на выбранной платформе Laravel была сформирована БД MySQL, которая хранит все собранные тексты в табличной форме.

Поиск текстов на чувашском языке для корпуса осуществлялся преимущественно из открытых источников в интернете. Были использованы художественные, публицистические, религиозные и другие тексты, их хронологические рамки — с 1906 по 2024 г.

На четвертом этапе перед вводом текстов в БД корпуса чувашского языка осуществлялась их обработка, которая включала в себя следующее: а) фильтрация входного текста через систему макросов для документов Microsoft Word; б) выявление нестандартных символов с помощью системы фильтров и последующая их замена. Остановимся более подробно на этих технических аспектах.

Обработка текстов в формате Microsoft Word через систему макросов помогает сделать данные единообразными и удобными для последующей работы. Нами были разработаны макросы для преобразования текстов, использующих шрифты Times New Roman Chuv и TimesET Chuvash, в стандартный шрифт Times New Roman. Работа макросов осуществляется следующим образом: первый макрос ищет в тексте чувашские символы (во всех чувашских шрифтах они представлены следующим образом: «/», «–», «+», «=» и т. д.) и заменяет их на комбинации букв и знаков, например, «Е_», «е_», «А_», «а_»). Данная процедура проводится только в тех частях текста, которые написаны шрифтом Times New Roman Chuv и TimesET Chuvash. Это необходимо, чтобы временно упростить представление текста и подготовить его для следующего этапа обработки. Затем второй макрос выполняет обратную операцию: он берет эти временные комбинации букв, например, «Е_», «е_», «А_», «а_» и заменяет их на правильные чувашские буквы, такие как «Ĕ», «ĕ», «Ă», «ă». Если первый макрос помогает найти и отметить нужные символы, то последующий возвращает им правильный вид — в результате данной операции в тексте отображаются все чувашские символы в нужной кодировке.

Далее идет обработка текста с помощью системы фильтров с целью выявления нестандартных символов, т. е. символов, отсутствующих в вышеприведен-

ном массиве S1. Если не провести данную операцию, то при поиске какого-либо слова на сайте корпуса, введенного с кодировкой, не соответствующей нами принятому Unicode, мы не получим результата поиска. Опишем пошагово, как осуществляется алгоритм фильтрации и выявления нестандартных символов при нажатии кнопки «Поиск юникода JS (рис. 1).

Рис. 1. Выявление нестандартных символов с помощью фильтров

Шаг 1. Начало работы.

Шаг 2. Задание стандартного набора символов. На данном этапе программа определяет массив S1, содержащий стандартные символы.

Шаг 3. Анализ текста. Пользователь вводит текст для анализа в текстовое поле textarea2 (поле «Входной текст для обработки»). Происходит запуск JavaScript-функции, которая обрабатывает введенный текст по следующему алгоритму:

1) текст считывается из текстового поля textarea2 и сохраняется в переменную content;

2) проводится итерация по символам текста: каждый символ текста проверяется, содержится ли он в массиве S1, если символ отсутствует в массиве S1, он считается нестандартным и записывается в массив result в формате <символ> — <Юникод-кодировка>;

3) устранение повторов: чтобы избежать дублирования записей, используется функция Set, которая удаляет повторяющиеся элементы из массива result.

Шаг 4. Вывод результата. После завершения обработки текстовое поле textarea1 (поле «Символы и их юникод кодировка») заполняется списком нестандартных символов и их юникод-кодировкой, разделенных запятыми. Если все символы в тексте соответствуют стандарту, поле остается пустым.

Шаг 5. Завершение работы.

В результате работы данного алгоритма выявляются все символы, несоответствующие стандарту массива S1. Степень выявления нестандартных символов довольно высокая, так как в текстах часто вместо одного символа с одним юникодом используется множество его вариаций. Например, для тире и дефиса в Unicode существует несколько различных символов: hyphen — U+002D, en dash — U+2013, em dash — U+2014, figure dash — U+2012, horizontal bar — U+2015, soft hyphen — U+00AD.

Алгоритм обработки текстов и замены нестандартных символов при нажатии кнопки «Преобразовать текст» происходит следующим образом:

Шаг 1. Начало работы.

Шаг 2. В переменную content1 (поле «Введите текст для преобразования») записывается основной текст, который нужно обработать, а в переменную $sin вносятся символы из поля UnicodeC (поле «Введите символы, использованные в тексте…»).

Шаг 3. Инициализация вспомогательных массивов $s1, $s2, где $s1 — массив для хранения заменяемых символов, $s2 — массив для хранения заменяющих значений для символов из $s1.

Далее выявленные нестандартные символы корректируются с помощью системы фильтров (рис. 2).

Рис 2. Коррекция нестандартных символов с помощью фильтров

Шаг 4. Выполняется цикл, который проходит по каждому символу строки $sin с учетом кодировки UTF-8 и извлекает текущий символ с помощью функции mb_substr. Затем с помощью конструкции switch проверяется, соответствует ли символ одному из заданных условий (если этот символ из нестандартных, то для него определена замена в массиве $s2). Если совпадение найдено, оригинальный символ добавляется в массив $s1, а его замена записывается в массив $s2.

В результате работы цикла формируются два массива: $s1 содержит символы из строки $sin, которые нужно заменить, а $s2 — соответствующие заменители. Эти массивы затем используются для замены символов в другой строке.

Шаг 5. Производится замена символов в строке $content1 с использованием массивов $s1 и $s2, которые были сформированы в цикле. Функция str_replace заменяет все вхождения символов из массива $s1 на соответствующие символы из массива $s2. Это позволяет преобразовать исходный текст, содержащий нестандартные или нежелательные символы, в текст с корректными заменителями, заданными в кодировке Unicode. После выполнения замены строка $string содержит обновленный текст, где все указанные символы были заменены.

Шаг 6. Возврат результата. Обработанный текст передается на отображение в представлении FR.index с параметром string (в поле «Результаты обработки»).

Шаг 7. Завершение работы.

Примеры замен:

Ă → Ӑ: нестандартный символ «Ă» с Unicode U+0102 заменяется на символ «Ӑ» с Unicode U+04D0 из массива S1;

ă → ӑ: нестандартный символ «ă» с Unicode U+0103 заменяется на «ӑ» с Unicode U+04D1 из массива S1.

В результате проведения такой обработки все тексты были унифицированы и подготовлены к дальнейшему вводу их в БД и отображению на веб-сайте корпуса.

На пятом этапе создания корпуса чувашского языка была разработана панель администрирования и осуществлен ввод обработанных текстов в БД MySQL, подключенную к платформе Laravel. В панель администрирования была добавлена форма заполнения БД корпуса (рис. 3).

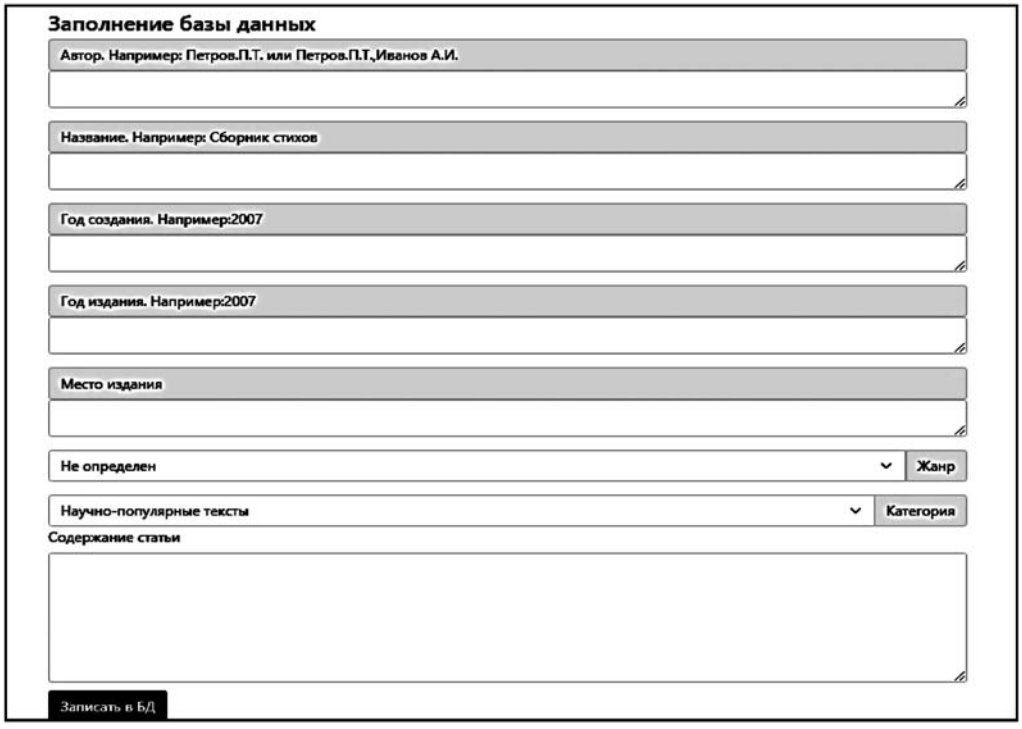


Рис. 3. Форма заполнения базы данных на панели администрирования Национального корпуса чувашского языка

Таким образом, нами были разработаны инструменты для фильтрации и преобразования текстов, а также формы для удобного внесения информации в БД корпуса. Ввод данных был стандартизирован, что повышает качество корпуса.

После введения обработанных текстов в БД был запущен в интернете сайт Национального корпуса чувашского языка, он доступен с июля 2024 г. по адресу — https://chuvkorpus.ru.

Интерфейс сайта корпуса чувашского языка удобный и интуитивно понятный (рис. 4). Он разработан на фреймворке Bootstrap 5, предоставляющем обширный набор инструментов для построения современного веб-дизайна. Для интерфейса сайта корпуса характерны следующие черты: 1) адаптивность, обеспечивающая доступ к сайту корпуса с любых устройств благодаря системе сеток; 2) гибкость, позволяющая быстро настраивать и добавлять на сайте новые элементы; 3) эргономичность, облегчающая интернет-пользователям доступ к нужным функциям. На сайте корпуса запущена навигационная панель, настроена форма поиска для быстрого ввода и отправки параметров запроса, визуализованы данные с помощью таблиц и карточек для структурированного представления результатов поиска.

Рис. 4. Сайт Национального корпуса чувашского языка

В центральной части сайта корпуса чувашского языка можно увидеть раздел под названием «Главная», где размещены вкладка «Информация о базе данных» и поисковая система, а также разделы «Категория» и «Жанр», которые, в свою очередь, делятся на несколько подразделов. Расположенные на сайте корпуса вкладки «Вход» и «Регистрация» предназначены для доступа к панели администрирования. Корректоры или системные администраторы проходят процедуру регистрации, после чего им присваиваются уровни доступа к БД корпуса.

БД корпуса чувашского языка постоянно развивается и пополняется. На начало 2025 г. она включала в себя 7,4 млн слов и 854,7 тыс. предложений. Весь массив текстов корпуса разбит на категории и жанры. Так, тексты распределены по 8 различным категориям. Наиболее многочисленные категории формируют прозаические, поэтические и публицистические тексты. Количественный состав текстов корпуса по категориям представлен в таблице 1.

Таблица 1

Количественный состав текстов Национального корпуса чувашского языка по категориям (БД на начало 2025 г.)

№	Категории	Число текстов
1	Драматургия	106
2	Научно-популярные тексты	10
3	Поэтические тексты	4249
4	Прозаические тексты	4291
5	Публицистические тексты	4010
6	Словари	11
7	Устное народное творчество	213
8	Церковно-религиозные тексты	62

Тексты в корпусе размещены также по 33 разнообразным жанрам. Больше всего текстов представлено в таких жанрах, как стихотворение, роман, рассказ и повесть. Количественный состав текстов корпуса по жанрам приведен в таблице 2.

Таблица 2

Количественный состав текстов Национального корпуса чувашского языка по жанрам (БД на начало 2025 г.)

№	Жанры	Число текстов
1	Автобиография	2
2	Анекдот	1
3	Баллада	7
4	Басня	40
5	Быль	5
6	Драма	1
7	Зарисовка	1
8	Инсценировка	1
9	Комедия	39
10	Мемуарная и эпистолярная литература	26
11	Монолог	2
12	Новелла	37
13	Очерк	94
14	Памфлет	1
15	Песня	70
16	Повесть	992
17	Поэма	217
18	Предисловие	21
19	Пьеса	24
20	Рассказ	1292
21	Рецензия, отзыв	6
22	Роман	1561
23	Романс	1
24	Сказка	193
25	Сонет	6
26	Стихотворение	3949
27	Сценка	2
28	Трагедия	30
29	Трагикомедия	5
30	Фельетон	2
31	Частушка	14
32	Эссе	9
33	Юмореска	204

Для Национального корпуса чувашского языка была разработана расширенная поисковая система, позволяющая искать нужный материал по автору, названию, слову или выражению, году, месту, жанру и категории, что делает его удобным для исследователей (рис. 5). Корпус предоставляет возможность находить данные по сложным запросам.

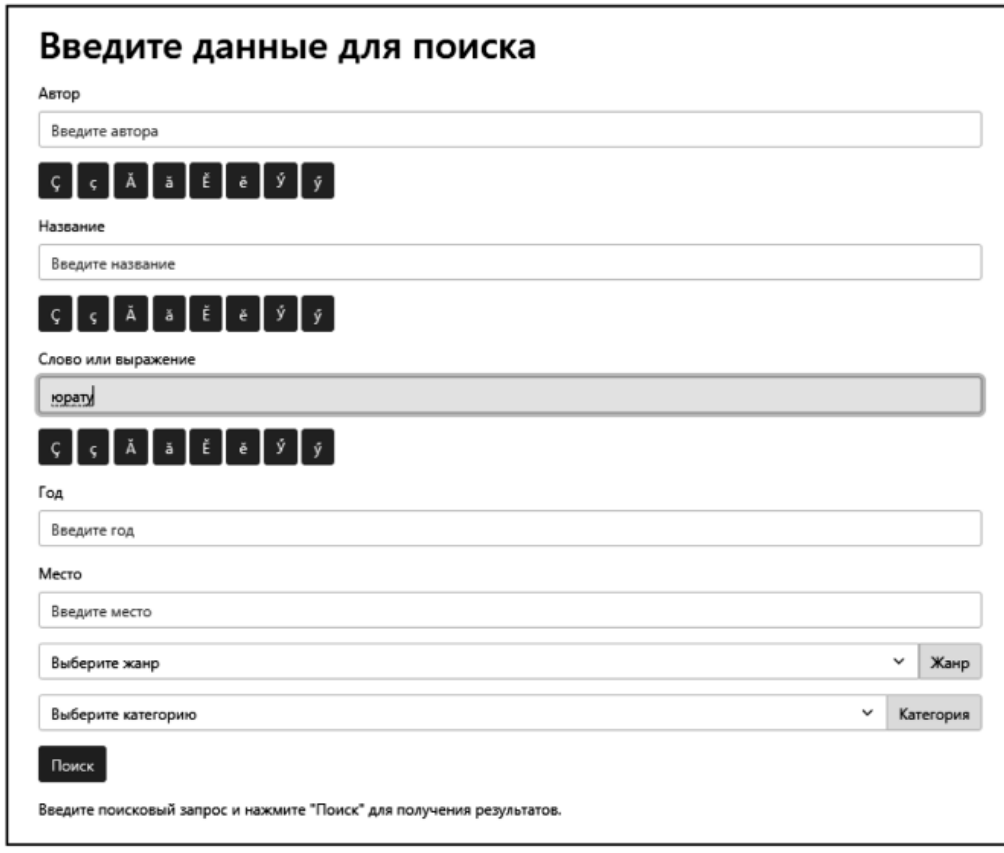


Рис. 5. Форма для расширенного поиска
на сайте Национального корпуса чувашского языка

Опишем пошагово алгоритм функции расширенного поиска.

Шаг 1. Начало работы.

Шаг 2. Извлекаем параметры запроса из объекта $request (он представлен такими полями, как «Автор», «Название», «Слово или выражение», «Год», «Место», «Выберите жанр», «Выберите категорию»).

Шаг 3. Создаем базовый запрос к модели MainTable, которая отвечает за выборку данных.

Шаг 4. Применяем фильтры. Если соответствующие параметры запроса не пустые, то они добавляются к запросу с помощью условий where [для полей «Автор», «Название», «Слово или выражение», «Год», «Место» применяется фильтр LIKE, а для полей «Выберите жанр» и «Выберите категорию» точное соответствие (=)].

Шаг 5. Пагинируем результаты запроса, т. е. сообщаем программе, какое количество результатов запроса выводить на одной странице (в нашем случае на одной странице отражается 100 записей).

Шаг 6. Извлекаем контекстные фрагменты. Если указан параметр «Слово или выражение», то выполняется следующая последовательность действий:

1) для каждого результата выборки извлекается поле content (контекст);

2) контекст конвертируется в кодировку UTF-8;

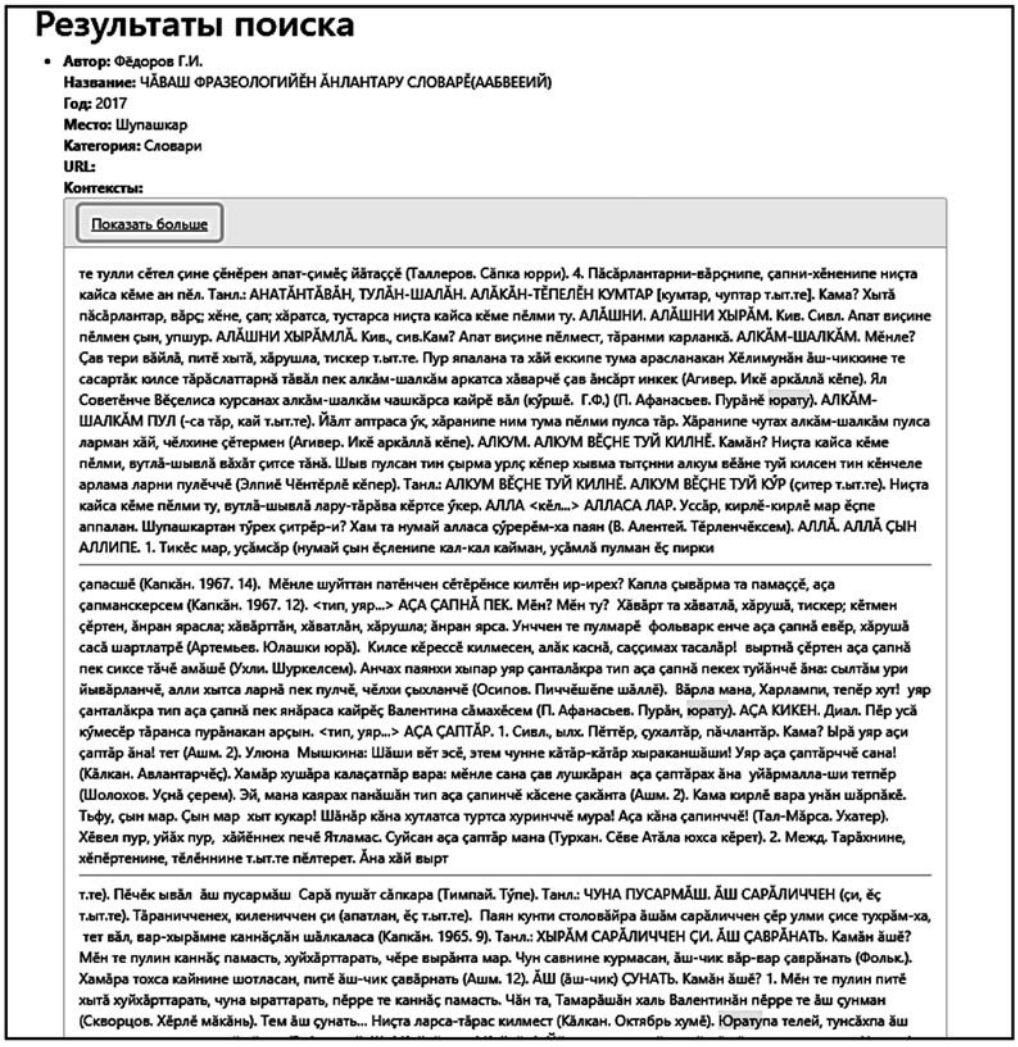
3) определяется позиция заданного слова в тексте; если слово найдено, вокруг него извлекается набор контекстов длиной 800 символов (до и после);

4) все найденные контексты сохраняются в массив $GLcontexts (глобальный контекст для всех слов) с привязкой к результату.

Шаг 7. Передаем в представление результаты поиска, заданное слово и контексты слова.

Шаг 8. Завершение работы.

Покажем на конкретном примере, как работает функция расширенного поиска на сайте корпуса (рис. 6).



Результаты поиска

- **Автор:** Фёдоров Г.И.
  Название: ЧĂВАШ ФРАЗЕОЛОГИЙĔН ĂНЛАНТАРУ СЛОВАРĔ(ААБВЕЕИЙ)
  Год: 2017
  Место: Шупашкар
  Категория: Словари
  URL:
  Контексты:

Показать больше

те тулли сĕтел сине çĕнĕрен апат-çимĕç йăтаççĕ (Таллеров. Сăпка юрри). 4. Пăсăрлантарни-вăрçнипе, çапни-хĕненипе ниçта кайса кĕме ан пĕл. Танл.: АНАТĂНТĂВĂН, ТУЛĂН-ШАЛĂН. АЛĂКĂН-ТĔПЕЛĔН КУМТАР [кумтар, чуптар т.ыт.те]. Кама? Хытă пăсăрлантар, вăрç; хĕне, çап; хăратса, тустарса ниçта кайса кĕме пĕлми ту. АЛĂШНИ. АЛĂШНИ ХЫРĂМ. Див. Сивл. Апат виçине пĕлмен çын, упшур. АЛĂШНИ ХЫРĂМЛĂ. Див., сив.Кам? Апат виçине пĕлмест, тăранми карланкă. АЛКĂМ-ШАЛКĂМ. Мĕнле? Çав тери вăйлă, питĕ хытă, хăрушла, тискер т.ыт.те. Пур япалана та хăй еккипе тума арасланакан Хĕлимунăн йăш-чиккине те сасартăк кулсе тăрăслаттарнă тăвăл пек алкăм-шалкăм арканса хăварчĕ çав ăнсăрт инкек (Агивер. Икĕ аркăллă кĕпе). Ял Советĕнче Вĕçелиса курсанах алкăм-шалкăм чашкăрса кайрĕ вăл (кӳршĕ. Г.Ф.) (П. Афанасьев. Пурăнĕ юрату). АЛКĂМ-ШАЛКĂМ ПУЛ (-са тăр, кай т.ыт.те). Йăлт аптраса ӳкĕ, хăранипе ним тума пĕлми пулса тăр. Хăранипе чутах алкăм-шалкăм пулса ларман хăй, чĕлхине çĕтермен (Агивер. Икĕ аркăллă кĕпе). АЛКУМ. АЛКУМ ВĔÇНЕ ТУЙ КИЛНĔ. Камăн? Ниçта кайса кĕме пĕлми, вутлă-шывлă вăхăт çитсе тăнă. Шыв пулсан тин çырма урлç кĕпер хывма тытçнни алкум вĕăне туй килсен тин кĕнчеле арлама ларни пулĕччĕ (Элпиĕ Чĕнтĕрлĕ кĕпер). Танл.: АЛКУМ ВĔÇНЕ ТУЙ КИЛНĔ. АЛКУМ ВĔÇНЕ ТУЙ КӲР (çитер т.ыт.те). Ниçта кайса кĕме пĕлми ту, вутлă-шывлă лару-тăрăва кĕртсе ӳкер. АЛЛА <кĕл...> АЛЛАСА ЛАР. Уссăр, кирлĕ-кирлĕ мар ĕçпе аппалан. Шупашкартан тӳрех çитрĕр-и? Хам та нумай алласа çӳрерĕм-ха паян (В. Алентей. Тĕрленчĕксем). АЛЛĂ. АЛЛĂ ÇЫН АЛЛИПЕ. 1. Тикĕс мар, уçăмсăр (нумай çын ĕçленипе кал-кал кайман, уçăмлă пулман ĕç пирки

çапасшĕ (Капкăн. 1967. 14). Мĕнле шуйттан патĕнчен сĕтĕрĕнсе килтĕн ир-ирех? Капла сывăрма та памаççĕ, аçа çапманскерсем (Капкăн. 1967. 12). <тип, уяр...> АÇА ÇАПНĂ ПЕК. Мĕн? Мĕн ту? Хăвăрт та хăватлă, хăрушă, тискер; кĕтмен çĕртен, ăнран ярасла; хăвăрттăн, хăватлăн, хăрушла; ăнран ярса. Унччен те пулмарĕ фольварк енче аçа çапнă евĕр, хăрушă сасă шартлатрĕ (Артемьев. Юлашки юрă). Килсе кĕрессĕ килмесен, алăк каснă, саççимах тасалăр! выртнă çĕртен аçа çапнă пек сиксе тăчĕ амăшĕ (Ухли. Шуркелсем). Анчах паянхи хыпар уяр çанталăкра тип аçа çапнă пекех туйăнчĕ ăна: сылтăм ури йывăрланчĕ, алли хытса ларнă пек пулчĕ, чĕлхи сыхланчĕ (Осипов. Пиччĕшĕпе шăллĕ). Вăрла мана, Харлампи, тепĕр хут! уяр çанталăкра тип аçа çапнă пек янăраса кайрĕç Валентина сăмахĕсем (П. Афанасьев. Пурăн, юрату). АÇА КИКЕН. Диал. Пĕр усă кӳмесĕр тăранса пурăнакан арçын. <тип, уяр...> АÇА ÇАПТĂР. 1. Сивл., ылх. Пĕттĕр, çухалтăр, пăчлантăр. Кама? Ырă уяр аçи çаптăр ăна! тет (Ашм. 2). Улюна Мышкина: Шăши вĕт эсĕ, этем чунне кăтăр-кăтăр хыраканшăши! Уяр аçа çаптăрччĕ сана! (Кăлкан. Авлантарчĕç). Хамăр хушăра калаçатпăр вара: мĕнле сана çав лушкăран аçа çаптăрах ăна уйăрмалла-ши тетпĕр (Шолохов. Уçнă çерем). Эй, мана каярах панăшăн тип аçа çапинчĕ кăсене çакăнта (Ашм. 2). Кама кирлĕ вара унăн шăрпăкĕ. Тьфу, çын мар. Çын мар хыт кукар! Шăнăр кăна хутлатса туртса хуринччĕ мура! Аçа кăна çапинччĕ! (Тал-Мăрса. Ухатер). Хĕвел пур, уйăх пур, хăйĕннех печĕ Ятламас. Суйсан аçа çаптăр мана (Турхан. Сĕве Атăла юхса кĕрет). 2. Межд. Тарăхнине, хĕпĕртенине, тĕлĕннине т.ыт.те пĕлтерет. Ăна хăй выр

т.те). Пĕчĕк ывăл ăш пусармăш Сарă пушăт сăпкара (Тимпай. Тӳпе). Танл.: ЧУНА ПУСАРМĂШ. ĂШ САРĂЛИЧЧЕН (çи, ĕç т.ыт.те). Тăраниччеченех, киленичченех çи (апатлан, ĕç т.ыт.те). Паян кунти столовăйра ăшăм сарăличчен çĕр улми çисе тухрăм-ха, тет вăл, вар-хырăмне каннăçлăн шăлкаласа (Капкăн. 1965. 9). Танл.: ХЫРĂМ САРĂЛИЧЧЕН ÇИ. ĂШ ÇАВРĂНАТЬ. Камăн ăшĕ? Мĕн те пулин каннăç памасть, хуйхăрттарать, чĕре вырăнта мар. Чун савнине курмасан, ăш-чик вăр-вар çаврăнать (Фольк.). Хамăра тохса кайнине шотласан, питĕ ăш-чик çаврăнать (Ашм. 12). ĂШ (ăш-чик) ÇУНАТЬ. Камăн ăшĕ? 1. Мĕн те пулин питĕ хытă хуйхăрттарать, чуна ыраттарать, пĕрре те каннăç памасть. Чăн та, Тамарăшăн халь Валентинăн пĕрре те ăш çунман (Скворцов. Хĕрлĕ мăкăнь). Тем ăш çунать... Ниçта ларса-тăрас килмест (Кăлкан. Октябрь хумĕ). Юратупа телей, тунсăхпа ăш

Рис. 6. Контекст вокруг найденного слова
в результатах расширенного поиска

При вводе слова, например, *юрату* «любовь» в поле «Слово или выражение» мы получим такие результаты поиска, в которых будут представлены все контексты из БД корпуса с найденным словом *юрату* (выйдут такие названия текстов, как «Чăваш фразеологийĕн ăнлантару словарĕ», «Библи Екклесиаст», «Библи Иезекииль» и т. д.). Каждую запись результатов поиска можно подробно рассмотреть, нажав на кнопку «Показать больше», в результате чего раскроется полный контекст найденного слова для данной записи.

Следует отметить, что все результаты поиска в корпусе чувашского языка ограничены минимальным контекстом, что позволяет получить интернет-пользователям достаточно данных для научного анализа или изучения языка, одновременно исключая возможность использования ими полных текстов без разрешения правообладателей. Такой подход гарантирует баланс между доступностью корпуса и правами авторов. При этом доступ к корпусу абсолютно бесплатный. Это позволяет каждому желающему использовать корпус для исследовательских или образовательных целей.

Обозначим дальнейшие перспективы развития электронного корпуса чувашского языка на базе платформы Laravel:

1) расширение БД корпуса за счет добавления новых текстов;

2) внедрение механизмов автоматической обработки текстов корпуса, прежде всего таких, как морфологическая разметка (т. е. для каждого слова в зависимости от принадлежности к той или иной части речи будут указаны его характеристики — число, падеж или время, лицо, основа, суффиксы) и синтаксическая разметка (исследователь сможет найти все словоформы, выступающие в определенной синтаксической роли);

3) улучшение поисковой системы корпуса, т. е. установка более совершенных алгоритмов поиска, например, полнотекстового поиска;

4) создание параллельного корпуса — разработка функционала для автоматического ввода параллельных текстов, а также поисковой системы для параллельного корпуса;

5) внедрение функционала для визуализации статистических данных корпуса, к примеру, добавление интерактивных графиков и диаграмм частотности слов, распределения частей речи и т. д.;

6) создание открытого API для взаимодействия с другими приложениями и сервисами в целях обогащения корпуса дополнительной информацией.

Заключение

Создание электронного корпуса чувашского языка представляет собой сложный и многоэтапный процесс. На данный момент реализованы ключевые этапы его формирования: 1) корпус разработан, управляется и анализируется на современной платформе Laravel, обеспечивающей гибкость, масштабируемость и адаптируемость под специфические задачи лингвистического анализа; 2) выполнена унификация всех текстов корпуса на основе юникод-стандарта; 3) разработаны макросы для обработки документов Microsoft Word; 4) разработаны алгоритмы фильтрации входных текстов; 5) созданы удобная панель администрирования и расширенная система поиска на сайте корпуса.

Национальный корпус чувашского языка в настоящее время содержит минимально необходимое количество текстов различных жанров и категорий для проведения лингвистических исследований. Он предоставляет возможность специалистам изучать грамматические, лексические и стилистические особенности

чувашского языка, анализировать его эволюцию и связи с другими тюркскими языками. На основе корпуса возможно создание автоматизированных словарей, грамматик и других языковых инструментов. В долгосрочной перспективе корпус должен стать уникальным инструментом для изучения и сохранения чувашского языка.

СПИСОК ИСТОЧНИКОВ

1. Авезов С. С., Маринина Ю. А. Электронные корпуса: инновационный подход к обучению перевода // Периодика. Журнал современной философии, социальных и гуманитарных наук. 2023. 16 марта. С. 7 — 13.

2. Губанов А. Р. Национальный корпус чувашского языка: создание лексического поисковика в системе Java // Актуальные вопросы истории и культуры чувашского народа: сборник статей / сост. и науч. ред. Н. Г. Ильина. Чебоксары : ЧГИГН, 2015. С. 130 — 145.

3. Желтов П. В. Национальный корпус чувашского языка: концепция и архитектура. Чебоксары : Изд-во Чуваш. ун-та, 2017. 160 с.

4. Желтов П. В. Создание национального корпуса чувашского языка: проблемы и перспективы // Современные проблемы науки и образования: [сайт]. URL: https://science-education.ru/ru/article/view?id=19046 (дата обращения: 10.03.2025).

5. Желтов П. В., Губанов А. Р., Желтов В. П. Морфологический стандарт национального корпуса чувашского языка // Современные проблемы науки и образования: [сайт]. URL: https://science-education.ru/ru/article/view?id=20578 (дата обращения: 10.03.2025).

6. Корпус чувашского языка: [сайт]. URL: https://ru.corpus.chv.su (дата обращения: 28.01.2025).

7. Антонов А. Н. Двуязычный корпус чувашского языка для машинного перевода // Электронная письменность народов Российской Федерации — 2021 & IWCLUL 2021: материалы Международной научно-практической конференции (Сыктывкар, 23 — 24 сентября 2021 г.) / отв. ред. А. Р. Эмексузян. Сыктывкар : Коми республикан. акад. гос. службы и управления, 2022. С. 18 — 22.

8. CQPweb: [сайт]. URL: https://cqpweb.lancs.ac.uk/ (дата обращения: 28.01.2025).

9. Laravel: [сайт]. URL: https://laravel.com (дата обращения: 28.01.2025).

10. Laurence Anthony's Website. AntConc: [сайт]. URL: https://www.laurenceanthony.net/software/antconc/ (дата обращения: 28.01.2025).

11. NoSketch Engine: [сайт]. URL: https://nlp.fi.muni.cz/trac/noske (дата обращения: 28.01.2025).

12. Sketch Engine: [сайт]. URL: https://www.sketchengine.eu/ (дата обращения: 28.01.2025).

13. Voyant Tools: [сайт]. URL: https://voyant-tools.org/ (дата обращения: 28.01.2025).

REFERENCES

1. Avezov S. S., Marinina Y. A. Electronic corpora: an innovative approach to teaching translation // Periodical Journal of Modern Philosophy, Social Sciences and Humanities. 2023. March 16:7 — 13. (In Russ.)

2. Gubanov A. R. The Chuvash national corpus language: creation a lexical search engine in Java // Actual issues of history and culture of the Chuvash people: a collection of articles / edited by N. G. Ilyina. Cheboksary;2015:130 — 145. (In Russ.)

3. Zheltov P. V. The Chuvash national corpus language: concept and architecture. Cheboksary;2017:160. (In Russ.)

4. Zheltov P. V. The Chuvash language national corpus creation: problems and prospects // Modern problems of science and education: [website]. URL: https://science-education.ru/ru/article/view?id=19046 (reference date: 10.03.2025). (In Russ.)

5. Zheltov P. V., Gubanov A. R., Zheltov V. P. Morphological standard of the Chuvash language national corpus // Modern problems of science and education: [website]. URL: https://science-education.ru/ru/article/view?id=20578 (reference date: 10.03.2025). (In Russ.)

6. The Chuvash language Corpus: [website]. URL: https://ru.corpus.chv.su (reference date: 28.01.2025). (In Russ.)

7. Antonov A. N. The Chuvash language bilingual corpus for machine translation // Electronic Written Language of the Peoples in the Russian Federation — 2021 & IWCLUL 2021: Proceedings of the International Scientific and Practical Conference (Syktyvkar, September 23 — 24, 2021) / edited by A. R. Emeksuzyan. Syktyvkar;2022:18 — 22. (In Russ.)

8. CQPweb: [website]. URL: https://cqpweb.lancs.ac.uk/ (reference date: 28.01.2025). (In Russ.)

9. Laravel: [сайт]. URL: https://laravel.com (reference date: 28.01.2025). (In Russ.)

10. Laurence Anthony's Website. AntConc: [сайт]. URL: https://www.laurenceanthony.net/software/antconc/ (reference date: 28.01.2025). (In Russ.)
11. NoSketch Engine: [сайт]. URL: https://nlp.fi.muni.cz/trac/noske (reference date: 28.01.2025). (In Russ.)
12. Sketch Engine: [сайт]. URL: https://www.sketchengine.eu/ (reference date: 28.01.2025). (In Russ.)
13. Voyant Tools: [сайт]. URL: https://voyant-tools.org/ (reference date: 28.01.2025). (In Russ.)

Статья поступила в редакцию 26.02.2025;
одобрена после рецензирования 01.03.2025; принята к публикации 02.03.2025
Информация об авторе:
Леонтьев Денис Михайлович, младший научный сотрудник
Чувашского государственного института гуманитарных наук
(428015, Россия, г. Чебоксары, Московский пр., 29, корп. 1),
ORCID: https://orcid.org/0000-0002-8798-8132, denis.lentev@mail.ru
Конфликт интересов: автор заявляет об отсутствии конфликта интересов.
Автор прочитал и одобрил окончательный вариант рукописи.

The article was received by the editorial board 26.02.2025;
approved after reviewing 01.03.2025; accepted for publication 02.03.2025
Information about the author:
Denis M. Leontiev, junior researcher, Chuvash State Institute of Humanities
(29/1 Moskovsky Ave., Cheboksary, 428015, Russia),
ORCID: https://orcid.org/0000-0002-8798-8132, denis.lentev@mail.ru
Conflict of interests: the author declares that thare is no conflict of interests.
The author has read and approved the final version of the manuscript.

Статья редакцие 26.02.2025 çитнĕ;
рецензиленĕ хыççăн 01.03.2025 ырланă; 02.03.2025 пичете йышăннă
Автор çинчен:
Леонтьев Денис Михайлович, Чăваш патшалăх гуманитари ăслăлăхĕсен
институчĕн ăслăлăх кĕçĕн ĕçтешĕ (428015, Раççей, Шупашкар хули,
Мускав пр., 29, 1 корп.), ORCID: https://orcid.org/0000-0002-8798-8132, denis.lentev@mail.ru
Пайталăх конфликчĕ: автор пайталăх конфликчĕ çуккине пĕлтерет.
Автор ал çырăвăн юлашки вариантне вуласа тухнă, ырланă.