МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра математической кибернетики и компьютерных наук

РЕАЛИЗАЦИЯ ПРИЛОЖЕНИЯ-КОНВЕРТЕРА РАЗМЕЧЕННОГО ТЕКСТА В XML ДЛЯ САРАТОВСКОГО ДИАЛЕКТОЛОГИЧЕСКОГО КОРПУСА

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 411 группы
направления 02.03.02 — Фундаментальная информатика и информационные технологии
факультета КНиИТ
Бастрыкина Максима Александровича

Научный руководитель
зав.каф.техн.пр., доцент,к.ф.-м.н._____ И. А. Батраева

Заведующий кафедрой
к. ф.-м. н., доцент _____ С. В. Миронов

Саратов 2025

ВВЕДЕНИЕ

Изучение русского языка, его диалектов и говоров является важной частью лингвистических исследований, направленных на сохранение языкового и культурного наследия. Создание электронных корпусов диалектной речи предоставляет лингвистам программное обеспечение для проведения масштабных исследований.

На кафедре теории, истории языка и прикладной лингвистики Филологического факультета Саратовского государственного университета имени Н. Г. Чернышевского продолжается работа по формированию Саратовского диалектологического корпуса. В процессе подготовки материалов для корпуса используются специфические методы лингвистической разметки, показывающие различные уровни анализа текста.

Актуальность темы представляет из себя автоматизацию процесса преобразования диалектных текстов с существующей специфической разметкой в машиночитаемый формат XML. Разработка специализированного программного обеспечения (приложения-конвертера) позволит значительно ускорить проверку текстов на ошибки, внесения необходимые исправления и изменения, обеспечить контроль как входных, так и выходных текстов. Ранее все эти задачи выполнялись вручную, что требовало значительных трудозатрат и времени. Автоматизация процессов позволит оптимизировать подготовку текстовых материалов для включения в Саратовский диалектологический корпус, свести к минимуму количество ошибок, связанных с ручной обработкой и обеспечить стандартизацию данных, необходимую для эффективного функционирования корпуса.

Целью дипломной работы является проектирование и реализация программного обеспечения (конвертера), предназначенного для автоматического преобразования размеченного диалектного текста из простого текстового формата с использованием специфической лингвистической разметки в машиночитаемый формат XML.

Работа выполнялась на основе технического задания, предоставленного кафедрой, где была поставлена задача разработать программное обеспечение.

В процессе подготовки программного приложения необходимо провести исследование формата входных текстов, содержащих специализированную разметку, применяемую в лингвистике, в частности при работе с диалектными

корпусами для последующей формализации правил преобразования, обеспечивающих корректное представление данных в машиночитаемом формате XML. Также, на этапе проектирования предусматривается разработка алгоритмов автоматической проверки входных данных, что позволяет своевременно обнаруживать распространённые ошибки, такие как несогласованные скобки, некорректные маркеры и недопустимые символы.

Программное обеспечение создавалось с учётом особенностей будущей эксплуатации в учебной среде преподавателями и студентами. Поэтому задача создать простой и интуитивно понятный графический интерфейс, обеспечивающий удобный выбор входного файла, запуск процессов проверки и конвертации, отображение результатов и выявленных ошибок, а также сохранение итогового XML.

Важным требованием со стороны кафедры является универсальность и портативность программного обеспечения, приложение должно легко запускаться на любом компьютере учебного заведения и устанавливаться без необходимости дополнительной настройки, достаточно простого копирования с внешнего носителя.

Практической значимостью работы будет являться создание готового к использованию программного обеспечения, которое может быть непосредственно применено сотрудниками кафедры теории, истории языка и прикладной лингвистики для эффективной подготовки текстовых материалов диалектологического корпуса, обеспечивая необходимую стандартизацию и структурированность данных для дальнейшей лингвистической работы в рамках корпусной системы.

Структура и объем работы. Для выполнения поставленных задач написана данная выпускная квалификационная работа, включающая в себя введение, теоретическую часть, которая обуславливает выбор технологий и их описание, практическую часть с описанием разработки алгоритмов и создания самого приложения, заключения, список использованных источников из 20 наименований и двух приложений: А описывает содержимое флеш-накопителя, Б содержит часть кода. Работа изложена на 53 страницах.

1 Содержание теоретической части

Основной задачей стоит разработка программного обеспечения, предназначенного для обработки текстовой информации, что существенно сократит время на выполнение задач, ранее выполнявшихся вручную и требовавших значительных трудозатрат, особенно в таких областях, как лингвистика и анализ данных. Перед началом разработки необходимо провести анализ существующих подходов к решению аналогичных задач, а также выбрать технологии, наиболее подходящие для их эффективной реализации.

Работа в этом направлении активно ведется различными научными центрами России. Параллельно с созданием диалектного подкорпуса в составе Национального корпуса русского языка (НКРЯ), который решает задачи представления диалекта как функциональной разновидности общенародного языка с акцентом на отличиях от литературной нормы, в Саратовском государственном университете им. Н.Г. Чернышевского была начата разработка Саратовского диалектного текстового корпуса.

Лингвистический корпус —— это крупное, структурированное и репрезентативное собрание текстов (как письменных, так и расшифровок устной речи), сформированное по определённым принципам и предназначенное для исследований языка. СарДК (Саратовский диалектологический корпус) -—- это специализированный корпус, включающий в себя тексты, отражающие диалектные особенности речи, характерные для территории Саратовской области. Основной задачей корпуса является сохранение, систематизация и анализ региональных вариантов русского языка. Работа по созданию СарДК ведется на кафедре теории, истории языка и прикладной лингвистики. Для сбора и первичной подготовки материалов корпуса используются собственные методы фиксации и разметки устной диалектной речи, разработанные работниками кафедры. Эта разметка имеет свою специфику и отличается от общих стандартов корпусной разметки, что обуславливает необходимость создания специального программного обеспечения для преобразования текстов из исходного формата в машиночитаемый формат, пригодный для включения в корпусную систему СарДК.

Активное формирование теоретических основ и начало работы над СарДК связаны с деятельностью ведущих ученых кафедры теории, истории языка и прикладной лингвистики. Важную роль в разработке концепции коммуникатив-

ной диалектологии и ее применении к созданию корпуса сыграл профессор В.Е. Гольдин. Под его руководством и при активном участии сотрудников кафедры, таких как профессор О.Ю. Крючкова, доцент А.П. Сдобнова и другие, были сформулированы принципы сбора материала, его систематизации и разметки. Профессор О.Ю. Крючкова стала руководителем проекта по созданию мультимедийного лингвокультурологического корпуса русских народных говоров «Русская диалектная речь», который является ключевой составляющей СарДК.

Современные корпуса являются незаменимым для лингвистов, поскольку позволяют изучать функционирование языка в естественных условиях, выявлять закономерности употребления лексических единиц и грамматических конструкций, а также отслеживать динамику языковых изменений. В отличие от произвольного набора текстов, лингвистический корпус включает метаинформацию о каждом тексте (автор, дата создания, жанр и другие метаданные) и, как правило, содержит разметку (аннотацию), описывающую различные уровни языковой структуры.

Разметка в корпусах может быть многоуровневой и охватывать различные аспекты языковой структуры. Среди основных типов разметки выделяют:

структурную разметку, которая отражает организацию текста и включает разбиение на заголовки, абзацы, предложения, а также, в случае устной речи, — на реплики говорящих;

морфологическую разметку, предусматривающую указание для каждого слова его леммы (начальной формы) и набора грамматических признаков, таких как часть речи, падеж, число, род и др.;

синтаксическую разметку, описывающую синтаксические связи между словами внутри предложения;

семантическую разметку, направленную на указание значений слов или семантических отношений между ними;

а также **разметку специфических языковых явлений**, таких как фонетические особенности в диалектной речи, невербальные сигналы в устных текстах или именованные сущности.

В процессе создания лингвистических корпусов, в частности диалектологических, этап разметки является одним из наиболее трудоемких. Часто для первичной обработки текстов используются программы автоматического морфологического анализа, подобные Mystem и Dialing. Они способны быстро опре-

делить начальную форму слова (лемму) и присвоить ему стандартный набор грамматических тегов, представляя результат в формализованном виде.

Однако, стандартная автоматическая разметка не всегда учитывает специфические особенности диалектной речи, вариативность форм, фонетические отклонения или более тонкие лингвистические категории, которые важны для глубокого анализа в рамках таких специализированных корпусов, как СарДК. По этой причине результаты автоматической обработки требуют существенной ручной доработки. В ходе этой ручной работы добавляются не только дополнительные уровни разметки, но и специфические для данного корпуса структурные маркеры и элементы.

Процесс ручной разметки и коррекции, несмотря на высокую квалификацию специалистов, неизбежно связан с риском внесения синтаксических ошибок в разметку. Наличие ошибок в исходных данных может сделать текст непригодным для автоматизированной обработки и загрузки в корпусную систему. Это подчеркивает важность этапа проверки размеченного текста до его преобразования в финальный машиночитаемый формат.

Для хранения и обмена размеченными корпусами часто используются стандартные форматы, такие как TEI (англ. Text Encoding Initiative — Инициатива по кодированию текстов), представляющий собой набор рекомендаций по представлению машиночитаемых текстов, особенно в области гуманитарных наук, и основанный на языке разметки XML (англ. eXtensible Markup Language — расширяемый язык разметки). Использование стандартизированных или строго формализованных машиночитаемых форматов, таких как XML, имеет критически важное значение для последующей автоматизированной обработки, поиска и анализа корпусных данных с помощью специализированного программного обеспечения.

Преобразование текстовых данных в структурированный машиночитаемый формат является типовой задачей в области обработки естественного языка и информационных технологий. Этот процесс включает несколько ключевых этапов.

Чтение данных. На этом этапе происходит извлечение содержимого исходного текстового файла. Особое внимание следует уделить корректной обработке кодировки, чтобы обеспечить правильное отображение всех символов, включая специальные лингвистические обозначения и символы национальных

алфавитов.

Анализ структуры текста. На данном этапе производится извлечение структурных и содержательных элементов из линейной последовательности символов в соответствии с правилами применяемой разметки. В случае разметки, основанной на специальных символах и конструкциях, извлечение информации включает распознавание маркеров, определение границ элементов (начало и конец комментария, область действия тегов), а также извлечение атрибутивной информации, таких как грамматические признаки слов или номера тем.

Проверка корректности разметки. Этот этап направлен на выявление ошибок, допущенных при ручной разметке текста, таких как незакрытые скобки, некорректное расположение маркеров или опечатки в синтаксисе обозначений.

Для хранения и обмена размеченными корпусами часто используются стандартные форматы, такие как TEI (англ. Text Encoding Initiative —– Инициатива по кодированию текстов), представляющий собой набор рекомендаций по представлению машиночитаемых текстов, особенно в области гуманитарных наук, и основанный на языке разметки XML. Формат XML был предложен в 2007 году и в дальнейшем стал стандартом для многих лингвистических корпусов, в том числе Национального корпуса русского языка, который разрабатывается компанией Яндекс. Использование стандартизированных или строго формализованных машиночитаемых форматов, таких как XML, имеет критически важное значение для последующей автоматизированной обработки, поиска и анализа корпусных данных с помощью специализированного программного обеспечения.

Преобразование данных. На данном этапе распознанные элементы входной разметки и извлечённая информация преобразуются в машиночитаемый формат XML. Сведения о морфологическом разборе слова, представленные во входном файле в виде строки с тегами, преобразуются в набор XML атрибутов или вложенных элементов внутри тега, описывающего слово. При необходимости осуществляется дополнительная обработка, включающая использование словарей для унификации тегов, нормализации атрибутивных значений и приведения информации к единообразному виду, соответствующему требованиям корпусной системы.

На заключительном этапе осуществляется формирование и сохранение

данных в выходной файл с соблюдением всех требований, предъявляемых к структуре и синтаксису формата XML. Это включает корректное использование тегов и атрибутов, соблюдение иерархии элементов, а также обеспечение валидности документа в соответствии с заданной схемой или спецификацией.

Реализация программного обеспечения для решения задачи проверки и конвертации текстов с графическим интерфейсом пользователя может быть осуществлена с использованием различных технологий. Выбор конкретной технологии определяется требованиями технического задания, доступными ресурсами, и предполагаемой средой эксплуатации приложения.

Задача создания программного обеспечение для нужд кафедры теории, истории языка и прикладной лингвистики предъявляла ряд специфических требований, изложенных в техническом задании. В первую очередь главной задачей стоит точная обработка специфической лингвистической разметки и преобразование её в строго определённую XML структуру. Кроме того, требуется простой и понятный графический интерфейс, удобный для преподавателей и студентов, которые не являются специалистами в области информационных технологий. Также важным условием является портативность и легкость установки, приложение должно запускаться на стандартных компьютерах под управлением Windows в условиях университета, даже без подключения к сети Интернет, путем простого копирования файлов с внешнего носителя.

После рассмотрения различных технологий был сделан выбор в пользу C#, .NET Framework и Windows Forms. Это выбор стал хорошим сочетанием условий для выполнения поставленной задачи с учетом всех требований технического задания кафедры. Для анализа структуры текста и преобразования формата язык C# и стандартные библиотеки .NET предоставляют все необходимые технологии для работы со строками, символами, регулярными выражениями и файловым вводом-выводом.

Создание графического интерфейса пользователя с помощью Windows Forms будет реализован в виде начального каркаса, который будет доработан до более удобного и практичного пользования, позволит размещать и настраивать основные элементы управления, предназначенных для широкого круга пользователей с разным уровнем подготовки. Внешний вид интерфейса по умолчанию соответствует стандартному стилю операционных систем начала 2000-х годов.

Такой внешний вид выглядит устаревшим по современным требованиям, поэтому в проекте будут предусмотрены дополнительные стилистические доработки интерфейса, изменения шрифтов, отступов, цветовой гаммы, направленные на повышение визуального качества и удобства работы с программой.

Требование к портативности и легкости запуска будет достигнуто данной технологией в условиях операционной системы Windows, которая является стандартной для компьютеров университета. Приложение, собранное под распространённую версию .NET Framework, часто может запускаться путём простого копирования исполняемого файла на целевую машину. Это исключает необходимость установки дополнительных виртуальных машин или крупных библиотек, что делает установку с внешнего носителя максимально удобным и простым.

Выбор веб-приложения на базе Java/Spring не подошёл из-за требований к автономности и простоте локального развертывания без зависимости от серверной инфраструктуры и наличия Интернета. Использование C++ и Qt оказалось не уместным для решения задачи, где основная сложность связана с логикой обработкой специфической разметки, а не с требованиями к максимальной производительности или сложной графике. Кроме того, установка приложений на C++/Qt может быть менее простой в стандартной среде Windows, чем для .NET/Windows Forms.

Таким образом, технология C#, .NET Framework и Windows Forms быдет наилучшим выбором, обеспечивающим возможности обработки текста, удобством разработки графического интерфейса и полным соответствием требованиям кафедры к портативности и простоте работы приложения в университетских условиях. Кроме того, основная система СарДК реализована на языке C#, как и большинство компонентов, входящих в структуру. Это обеспечивает совместимость и технологическую взаимосвязь, а также упрощает потенциальную интеграцию разрабатываемого приложения в уже существующую систему кафедры.

2 Содержание практической части

Пользовательский интерфейс разработан с учетом удобства использования для лингвистов. Главное окно содержит два текстовых поля: левое - отображает исходного текста, правое - получаемый текст и набор кнопок для управления функциями. Основные функции, доступные через интерфейс: выбор входного файла (.txt), автоматическая проверка разметки после загрузки, ручная проверка текста в любое время, навигация по найденным ошибкам в исходном тексте, конвертация текста целиком или по абзацам, генерация "читаемого"текста без разметки, сохранение содержимого любого текстового поля в файл (.txt или .xml), а также специальная функция обработки файла для автоматического заполнения пропущенных элементов в разметке слова. Внешний вид основного окна приложения представлен на рисунке 1.

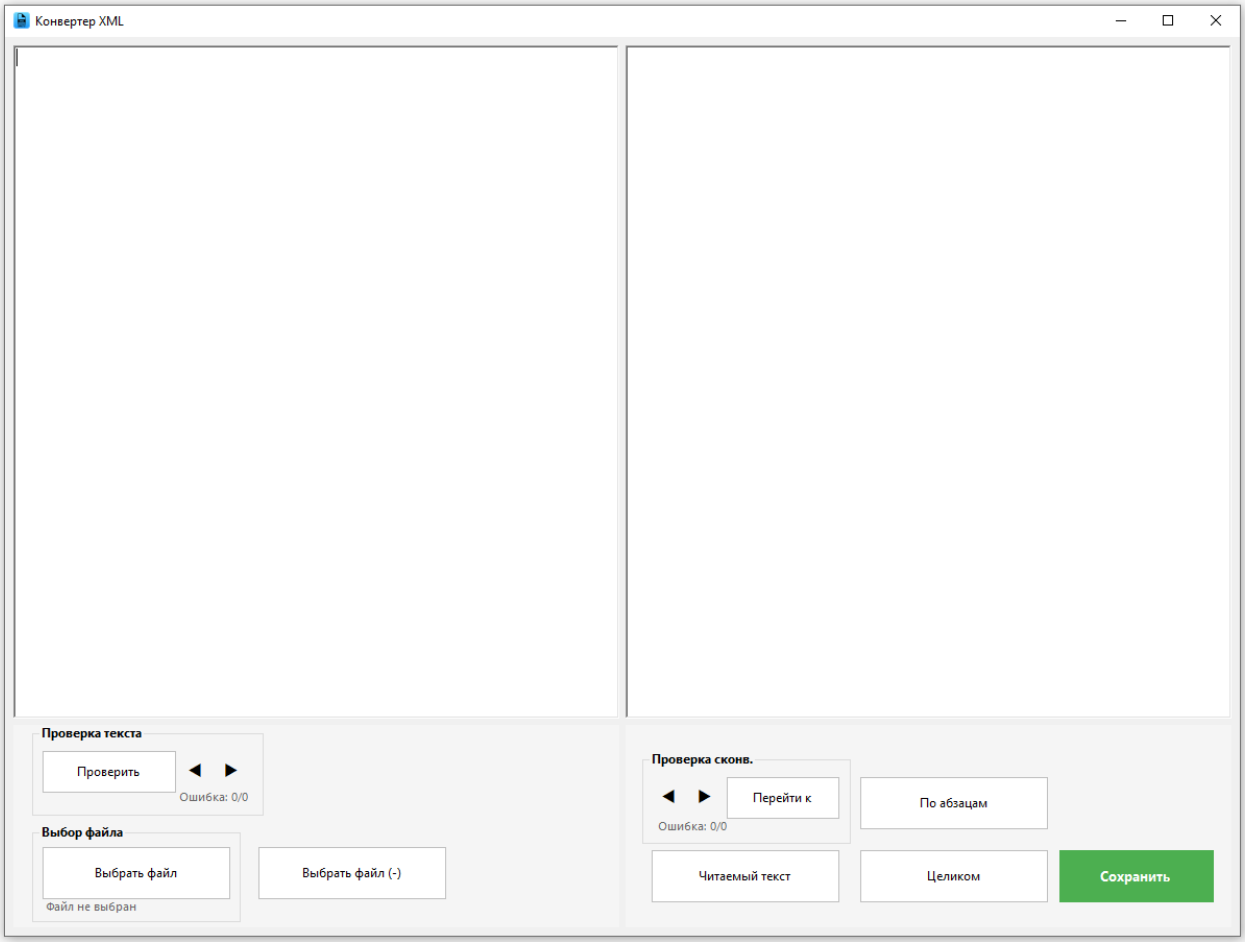

Рисунок 1 – Основное окно программного приложения

Приложение поддерживает работу с внешними файлами словарей, расположенными в подпапке рядом с исполняемым файлом. Эти файлы считывают-

ся в память при запуске и используются модулем логики для интерпретации маркеров темы/жанра и нормализации морфологических тегов. Такой подход позволяет обновлять словари без изменения кода приложения. Для удобства пользователей реализованы функции поиска и замены текста в текстовых полях.

Реализация логики обработки текста сосредоточена в классе Logic. Методы loadThemes и loadTags отвечают за загрузку словарей. Метод replacingPreviousWord реализует автоматическое заполнение пропущенных элементов в разметке слова с использованием регулярных выражений. Методы checkingMarkup, checkingBrackets, checkingPercentPairs реализуют алгоритмы проверки разметки на основе стеков, фиксируя ошибки с указанием позиции. Основной алгоритм конвертации реализован в методе convert и вспомогательных методах, которые последовательно обрабатывают текст, распознают элементы разметки, используют загруженные словари и алгоритм КМП для преобразования тегов, и формируют выходной XML документ в объекте stringBuilder.

Тестирование разработанного приложения проводилось поэтапно. Выполнялось модульное тестирование ключевых алгоритмов обработки на синтетических примерах разметки. Проводилось интеграционное тестирование взаимодействия между интерфейсом и логикой. Функциональное тестирование осуществлялось на реальных диалектных текстах, предоставленных кафедрой. Критерием успешности являлось получение синтаксически корректного XML, точно отражающего исходную разметку, правильное использование словарей и эффективное выявление ошибок. Тестирование подтвердило соответствие приложения требованиям и его готовность к использованию.

ЗАКЛЮЧЕНИЕ

С целью автоматизации процесса преобразования диалектных текстов в машиночитаемый формат XML, разработано программное обеспечение. Данный конвертер позволяет ускорить проверку, внести исправления и изменения, обеспечивает контроль как входных, так и выходных текстов. Автоматизация процессов позволяет оптимизировать подготовку текстовых материалов для включения в базу для эффективного функционирования Саратовского диалектологического корпуса.

Реализация программного обеспечения в рамках данной работы упростила преобразование размеченного диалектного текста из простого текстового формата с использованием специфической лингвистической разметки в машиночитаемый формат XML.

Программное обеспечение применимо для преподавателей и студентов. Учтена разработка простого и понятного интерфейса, удобного запуска процессов проверки и конвертации, а также отображения результатов и сохранения итогового XML.

В процессе реализации проекта проведено исследование формата входных текстов, включающих в себя специализированную разметку, используемую лингвистами при работе с диалектными материалами. На основе этого анализа были формализованы правила преобразования элементов входной разметки в соответствующие XML элементы и атрибуты. Разработаны алгоритмы проверки корректности входных данных, позволяющие выявлять типовые ошибки.

Результатом бакалаврской работы является разработанное программное приложение, полностью соответствующее поставленным требованиям и обеспечивающее эффективную автоматизацию процесса подготовки диалектных текстов к включению в корпус. Созданный конвертер успешно прошел апробацию на базе СарДК и в настоящее время используется сотрудниками кафедры теории, истории языка и прикладной лингвистики для обработки текстовых материалов. Полученное техническое задание, предоставленное кафедрой, выполнено.