Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н. Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н. Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К КУРСОВОЙ РАБОТЕ

НА ТЕМУ:

Разработка базы данных для АРМ разметчика параллельного корпуса технических текстов.

Студент ИУ7-64Б _____ К. А. Рунов
(Группа) (Подпись, дата) (И. О. Фамилия)

Руководитель курсовой работы _____ Ю. В. Строганов
(Подпись, дата) (И. О. Фамилия)

2024 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В современном мире немаловажное значение имеет корпусная лингвистика. Корпуса текстов находят применение в различных областях — в машинном переводе, в разработке словарей, в лингвистических исследованиях. Для того, чтобы из корпуса текстов можно было извлекать какую-то пользу, тексты в нем должны быть размечены. Существуют алгоритмы, позволяющие автоматически производить разметку. Но зачастую, для проверки ее корректности, все равно требуется вмешательство человека.

На данный момент не существует открытых параллельных корпусов технических текстов. Также нет открытых информационных систем, позволяющих одновременно

— производить разметку текста в параллельном корпусе,

— производить поиск по параллельному корпусу,

— организовать удобную работу множества разметчиков.

Создание такой информационной системы позволит во многом автоматизировать рабочее место разметчиков параллельного корпуса.

Целью данной работы является разработка базы данных для автоматизации рабочего места разметчиков параллельного корпуса технических текстов.

Задачи курсового проекта:

— провести анализ предметной области параллельных корпусов текстов;

— спроектировать сущности базы данных и ограничения целостности АРМ разметчика корпуса технических текстов;

— выбрать средства реализации базы данных и приложения;

— разработать сущности базы данных и реализовать ограничения целостности базы данных;

— описать интерфейс доступа к базе данных;

— исследовать зависимость времени ответа от количества запросов в секунду.

1 Аналитический раздел

1.1 Анализ предметной области

Корпусная лингвистика — раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий. [1, с. 11]

1.1.1 Корпуса текстов

Под лингвистическим, или языковым, корпусом текстов понимается большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач. [1, с. 11]

В понятие «корпус текстов» входит также поисковая система, позволяющая производить поиск по корпусу, называемая *корпусным менеджером*. Поиск в корпусе позволяет по любому слову построить *конкорданс* — список всех употреблений данного слова к контексте со ссылками на источник. [1, с. 12]

Выделяют как минимум три типа корпусов текстов:

— корпусы первого типа — универсальные, отражающие все многообразие речевой деятельности;

— корпусы второго типа — специфичные, отражающие бытование некоторого языкового или культурного явления в общественной речевой практике, например корпус пословиц или корпус политических метафор в газетной речи;

— корпусы третьего типа — специфичные, создаваемые для решения специальной задачи, например для обучения, для задач социолингвистики, для отладки систем машинного перевода. [1, с. 12]

Виды корпусов текстов

В таблице 1 приведена классификация корпусов текстов по разным признакам.

Признак	**Типы корпусов**
Цель	Многоцелевые, специализированные
Параллельность	Параллельные, сопоставимые
Тип языковых данных	Письменные, устные (речевые), смешанные
«Литературность»	Литературные, диалектные, разговорные, терминологические, смешанные
Жанр	Литературные, фольклорные, драматургические, публицистические
Назначение	Исследовательские, иллюстративные
Динамичность	Динамические (мониторные), статические
Разметка	Размеченные, неразмеченные
Характер разметки	Морфологические, синтаксические, семантические, анафорические, просодические и т. д.
Доступность	Свободно доступные, коммерческие, закрытые
Объем текстов	Полнотекстовые, «фрагментнотекстовые»

Таблица 1 – Классификация корпусов [1, с. 57]

Корпус технических текстов, для которого будет разрабатываться база данных в настоящей работе, относится к корпусам третьего типа и является:

— специализированным,

— параллельным,

— многоязыковым,

— письменным,

— терминологическим,

— динамическим (постоянно будет пополняться),

— размеченным,

— свободно доступным,

— полнотекстовым.

Применение корпусов текстов

1.1.2 Тексты и разметки

Классификация текстов

Структура технических текстов

Виды текстовых разметок

1.2 Существующие решения

1.3 Формализация задачи

1.4 Формализация данных

1.5 Сущности базы данных

1.6 Формализация и описание пользователей

1.7 Сценарии использования

1.8 Анализ существующих баз данных

1.8.1 Выбор базы данных

1.9 Вывод

2 Конструкторский раздел

2.1 Проектирование базы данных

2.2 Описание сущностей

2.3 Описание ограничений целостности

2.4 Описание функций, процедур и триггеров

2.5 Описание ролевой модели

2.6 Вывод

3 Технологический раздел

3.1 Выбор средств реализации

3.2 Описание реализаций

3.2.1 Сущности базы данных

3.2.2 Ограничения целостности базы данных

3.2.3 Ролевая модель на уровне базы данных

3.2.4 Функции, процедуры и триггеры

3.2.5 Тестирование

3.2.6 Интерфейс доступа к базе данных

3.3 Вывод

4 Исследовательский раздел

4.1 Технические характеристики

4.2 Описание исследования

4.3 Проведение исследования

4.4 Вывод

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Захаров В., Богданова С. Корпусная лингвистика. ЛитРес, 2022. – URL: https://books.google.ru/books?id=HpTcDwAAQBAJ.