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 пользу, тексты в нем должны быть размечены. Существуют алгоритмы, позволяющие автоматически производить разметку, но 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

В данном разделе будет проведен анализ предметной области корпусов текстов.

1.1 Анализ предметной области

Корпусная лингвистика — это раздел компьютерной лингвистики, который занимается разработкой принципов построения и использования корпусов текстов с помощью компьютерных технологий. Лингвистические корпуса представляют собой структурированные массивы данных, которые используются для изучения языковых единиц в текстах. В рамках корпуса существует поисковая система, которая позволяет находить необходимые языковые единицы и примеры их употребления благодаря технологии текстовой разметки. В свою очередь разметка может быть ручной и автоматической. [1]

Виды корпусов текстов

В таблице 1 приведена классификация корпусов текстов по разным признакам.

Признак	**Типы корпусов**
Цель	Многоцелевые, специализированные
Параллельность	Параллельные, сопоставимые
Динамичность	Динамические (мониторные), статические
Разметка	Размеченные, неразмеченные
Характер разметки	Морфологические, синтаксические, семантические, анафорические, просодические и т. д.
Объем текстов	Полнотекстовые, «фрагментнотекстовые»

Таблица 1 – Классификация корпусов [2, с. 57]

Корпус технических текстов, для которого будет разрабатываться база данных в настоящей работе, является специализированным, параллельным, многоязыковым, динамическим (будет постоянно пополняться).

Параллельные корпусы

Параллельные корпусы — корпусы, представляющие собой множество текстов-оригиналов, написанных на каком-либо исходном языке, и текстов — переводов этих исходных текстов на один или несколько других языков [2, с. 61].

При подготовке параллельных корпусов и разработке программ для их обработки, требуется выровнять тексты — установить соответствие между фрагментами текста оригинала и текста перевода. Для решения этой задачи существуют различные методы автоматического выравнивания текстов по предложениям, грамматическим конструкциям, терминам, словам и словосочетаниям. [2, с. 61]

Ниже приведен пример выравнивания текстов на уровне предложений.

1	THE PLAY — for which Briony had designed the posters, programs and tickets, constructed the sales booth out of a folding screen tipped on its side, and lined the collection box in red crepe paper — was written by her in a two-day tempest of composition, causing her to miss a breakfast and a lunch.	Пьеса, для которой Брайони рисовала афиши, делала программки и билеты, сооружала из ширмы кассовую будку и обклеивала коробку для денежных сборов гофрированной красной бумагой, была написана ею за два дня в порыве вдохновения, заставлявшего ее забывать даже о еде.
2	When the preparations were complete, she had nothing to do but contemplate her finished draft and wait for the appearance of her cousins from the distant north.	Когда приготовления закончились, ей не оставалось ничего, кроме как созерцать свое творение и ждать появления кузенов и кузины, которые должны были прибыть с далекого севера.

Таблица 2 – Пример выравнивания текстов на уровне предложений [2, с. 62]

Проблемы определения границ терминов

При попытке проведения автоматического выравнивания на уровне терминов, возникает ряд проблем. Среди них выделяют [1]:

— неправильное определение границ терминов-словосочетаний, состоящих из двух и более слов и составных терминов;

— распознавание составных терминов и терминов-словосочетаний, состоящих из двух и более слов; в частности, распознавание лексической единицы как части составного термина или как свободной лексической единицы;

— определение лексической единицы как термина в зависимости от контекста и тематики текста, в котором данная лексическая единица употребляется;

— объемные списки терминов-кандидатов, которые необходимо проверять вручную, поскольку частота не является достаточным критерием для оценки того, является ли выделенное слово термином или нет.

Точность определения границ термина при автоматической разметки является одной из основных лингвистических задач, а отсутствие на сегодняшний день веб-платформ для автоматической разметки русскоязычных текстов делает актуальным разработку таковой.

1.1.1 Типы текстовых разметок

Существует множество типов текстовых разметок. Большинство современных корпусов относятся к корпусам морфологического или синтаксического вида [2, с. 56].

Далее подробнее будут рассмотрены структурная и семантическая разметки, поскольку они являются основными типами разметки в параллельном корпусе технических текстов.

Структурная разметка

Структурная разметка предназначена для выделения структурных элементов текста (том, книга, часть, глава, действие, сноска, ремарка, стих, а также: абзац, предложение, словоформа и текстоформа — таблица, формула и др.) [3]. В контексте учебно-научных текстов структурная разметка использована для выделения названия статьи, авторов, оглавления, предисловия, введения и т.д.

На рисунке 1 представлена структурная схема элементов учебно-научного текста.

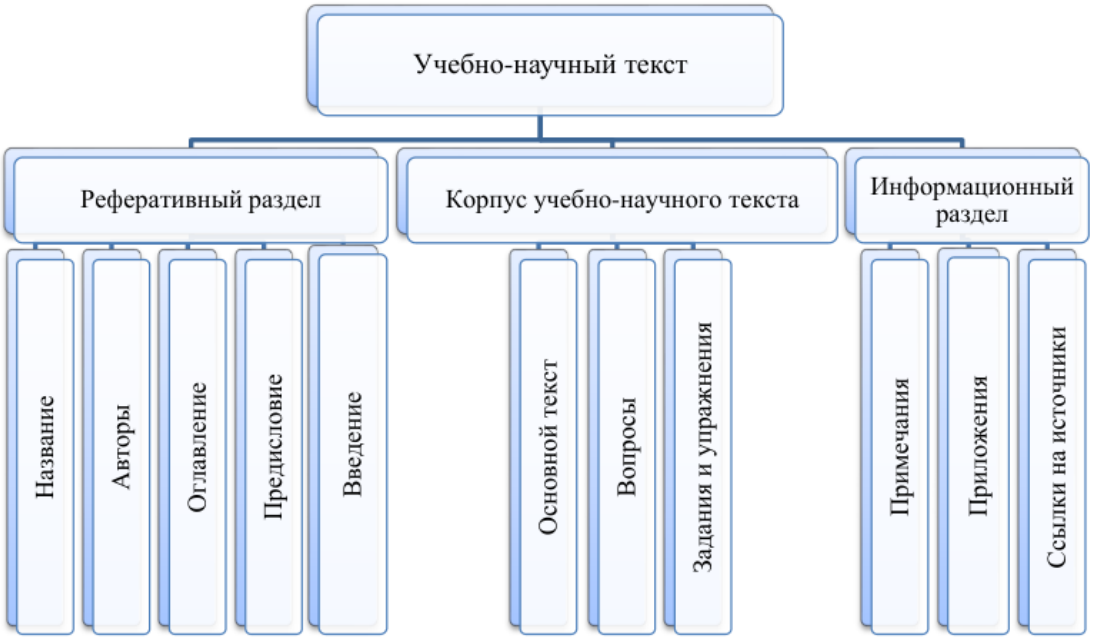

Рисунок 1 – Структурные элементы учебно-научных текстов [4]

Семантическая разметка

Семантическая разметка помогает установить контекст высказывания и устранить двусмысленность. Основная цель семантической разметки — «формализировать» значения слов и сделать тексты пригодными для машинной обработки. [5]

Существует множество видов семантических разметок. Один из вариантов русской семантической разметки представлен на сайте Центра компьютерных и корпусных языковых исследований Ланкастерского университета [6, 2, с. 51]:

A	Общие понятия
A1	Общие понятия
A1.1.1	Обычные действия / Изголовление ч.-л.
A1.1.1	Повреждение и разрушение
A1.2	Пригодность
A1.3	Осторожность
...	.....
B	Тело человека
B1	Анатомия и физиология
B2	Здоровье и болезнь
...	.....
C	Искусства и ремесла
...	.....

Таблица 3 – Русская семантическая разметка [6]

В научно-технических текстах для семантической разметки могут использоваться семантические падежи Ч. Филлмора [5].

1.2 Существующие решения

В современном мире существует множество параллельных корпусов (Opus, Linguee, MyMemory, Glosbe, Reverso, TAUS Data Cloud и др.) [7], сервисов для автоматического выравнивания текстов (Hunalign, Euclid, Abbyy Aligner, Trados, Winalign, Wordfast tools, Giza++ и др.) [2, с. 62], служб, позволяющих создавать собственные корпусы и производить в них поиск (SketchEngine [8], NoSketchEngine [9]); инструментов для автоматического извлечения терминов (TerMine, TermExtraction, Terminology Extraction) [1] и прочих инструментов для работы с корпусами текстов (OpenCorpora [10]).

Но на данный момент не существует открытых параллельных корпусов технических текстов [11]. Также нет открытых информационных систем, позволяющих одновременно производить разметку текста в параллельном корпусе, производить поиск по параллельному корпусу и организовать удобную работу множества разметчиков.

1.3 Формализация задачи

В ходе выполнения курсовой работы необходимо спроектировать и разработать базу данных для хранения документов — технических текстов на разных языках, их разметок, и информации о пользователях базы данных.

Для взаимодействия с базой данных необходимо разработать интерфейс, предоставляющий возможности

— добавления новых документов в базу данных,

— добавления новых разметок в базу данных,

— произведения поиска по текстам и разметкам, хранящимся в базе данных.

1.4 Формализация данных

Разрабатываемая база данных должна хранить информацию о следующих сущностях:

— пользователь;

— документ;

— метаданные о документе — автор;

— задание на разметку;

— структурная разметка;

— терминологическая разметка.

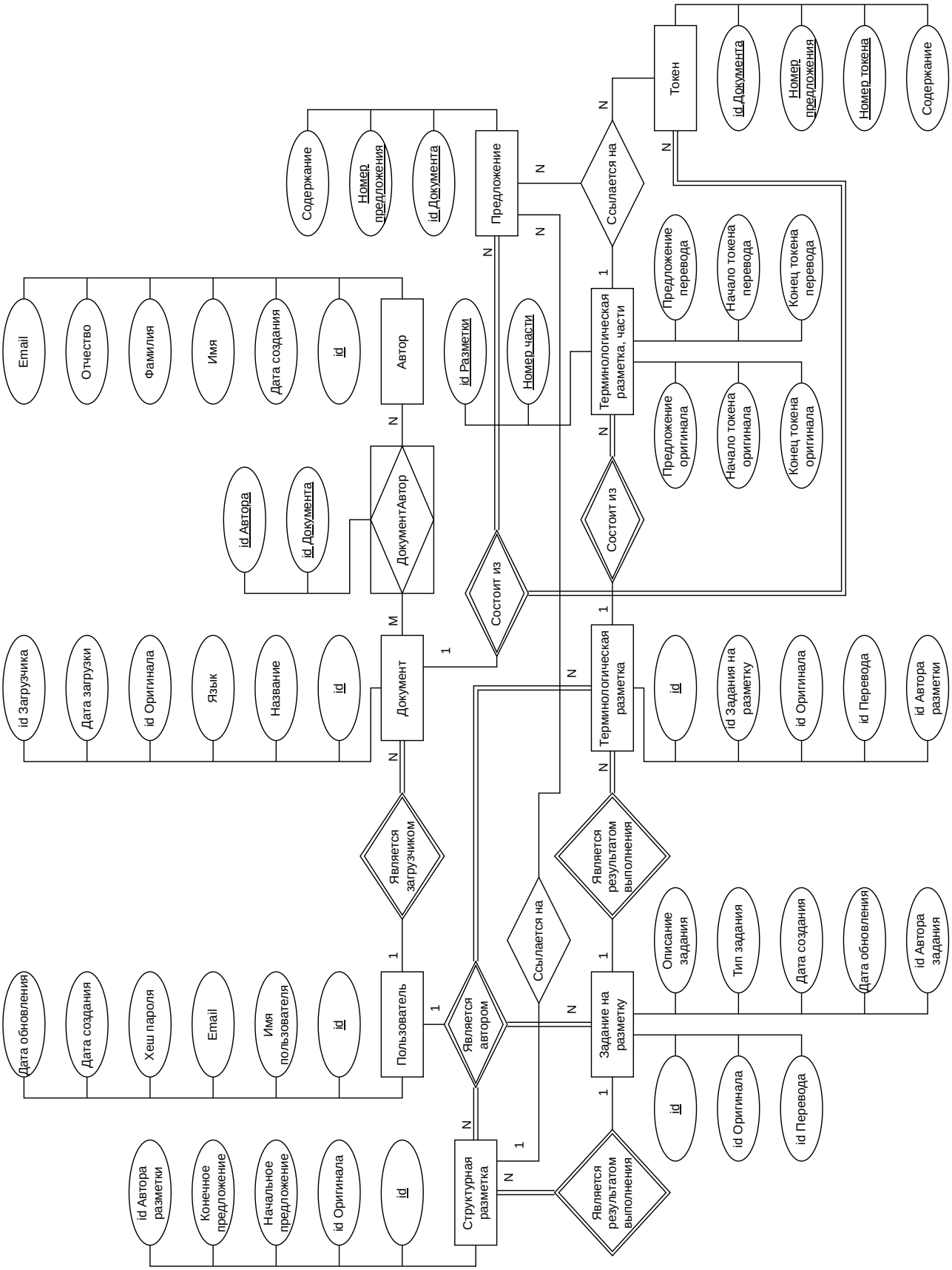

Рисунок 2 – ER-диаграмма в нотации Чена

1.5 Формализация и описание пользователей

1.6 Сценарии использования

1.7 Анализ существующих баз данных

1.7.1 Выбор базы данных

1.8 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. Бутенко Ю.И., Сулейманова Э.В. Терминологическая разметка научно-технических текстов в специальном корпусе // Проблемы лингвистики и лингводидактики в неязыковом вузе: 5-я Международная научно-практическая конференция. 2022. Т. 1. С. 329–337.

2. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник. 3-е изд., перераб. – СПб.: Изд-во С.-Петерб. ун-та, 2020. – 234 с.

3. Лесников В.С. Виды разметок текстовых корпусов русского языка // Научно-техническая информация. Сер. 2. Информационные процессы и системы. 2019. № 9. С. 27–30.

4. Бутенко Ю.И. Модель учебно-научного текста для разметки корпуса научно-технических текстов // Экономика. Информатика. 2021. Т. 48, № 1. С. 123–129.

5. Бутенко Ю.И., Попова Н.М. Особенности семантической разметки в корпусе научно-технических текстов // Проблемы лингвистики и лингводидактики в неязыковом вузе: 5-я Международная научно-практическая конференция. 2022. Т. 1. С. 324–328.

6. University Centre for Computer Corpus Research on Language [Электронный ресурс]. – URL: https://ucrel.lancs.ac.uk/usas (дата обращения: 20.05.2024).

7. Бутенко Ю.И., Киселёва А.Д. Анализ современных корпусов параллельных текстов // Актуальные проблемы лингвистики и лингводидактики в неязыковом вузе: 4-я Международная научно-практическая конференция. 2020. Т. 1. С. 238–242.

8. Sketch Engine: Create and search a text corpus [Электронный ресурс]. – URL: https://www.sketchengine.eu (дата обращения: 20.05.2024).

9. NoSketch Engine [Электронный ресурс]. – URL: https://nlp.fi.muni.cz/trac/noske (дата обращения: 20.05.2024).

10. OpenCorpora — открытый корпус [Электронный ресурс]. – URL: https://opencorpora.org (дата обращения: 25.03.2024).

11. Бутенко Ю.И., Строганов Ю.В., Бабаджанян Р.В. Исследовательский прототип параллельного корпуса научно-технических текстов // Актуальные проблемы лингвистики и лингводидактики в неязыковом вузе: 4-я Международная научно-практическая конференция. 2020. Т. 1. С. 205–209.