



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
(Группа)

К. А. Рунов
(Подпись, дата) (И. О. Фамилия)

Руководитель курсовой работы

Ю. В. Строганов
(Подпись, дата) (И. О. Фамилия)

2024 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	4
1 Аналитическая часть	5
1.1 Корпуса текстов	5
1.1.1 Виды	5
1.1.2 Применение	6
1.1.3 Устройство	8
1.2 Тексты	9
1.2.1 Виды разметок	9
1.2.2 Технические тексты и их структура	10
1.2.3 Проблема терминов	12
1.3 Существующие параллельные корпуса	12
1.4 Вывод	12
2 Конструкторская часть	13
2.1 Вывод	15
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	16
ПРИЛОЖЕНИЕ А	17

ВВЕДЕНИЕ

1 Аналитическая часть

В данной части будет идти речь о корпусах текстов, видах текстов и текстовых разметок.

1.1 Корпуса текстов

Корпусная лингвистика — раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий. 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. 5]

1.1.1 Виды

В таблице 1 ниже представлена классификация корпусов по некоторым признакам.

Таблица 1 – Классификация корпусов [1, с. 16]

Признак	Тип корпуса
Цель	многоцелевые, специализированные
Параллельность	параллельные, сопоставимые (псевдопараллельные)
Тип языковых данных	письменные, устные (речевые), смешанные
«Литературность»	литературные, диалектные, разговорные, терминологические, смешанные
Жанр	литературные, фольклорные, драматургические, публицистические
Назначение	исследовательские, иллюстративные
Разметка	размеченные, неразмеченные
Характер разметки	морфологические, синтаксические, семантические, анафорические и т. д.
Объем текстов	полнотекстовые, «фрагментнотекстовые»

В данной работе особое внимание уделяется параллельным корпусам, так как именно для работы с таким видом корпусов будет разрабатываться база данных.

Параллельный корпус — это двуязычный корпус. В нем хранится два множества текстов — оригиналов и их переводов. Работа с корпусом (выравнивание, разметка, поиск) производится сразу с двумя текстами.

Для возможности использования параллельного корпуса в качестве инструмента исследования, тексты должны быть выровнены — отдельные фрагменты оригинала должны совпадать с соответствующими фрагментами перевода [2].

Подробнее о применении, устройстве и проблематике параллельных корпусов будет говориться в следующих подразделах.

1.1.2 Применение

В данном подразделе рассматриваются некоторые применения параллельных корпусов.

Машинный перевод

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Обработка естественного языка (NLP)

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices

augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Кросс-языковой поиск информации (CLIR)

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Разработка словарей и учебных материалов

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Лингвистические исследования

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

1.1.3 Устройство

Параллельные корпуса устроены таким образом, чтобы решались поставленные перед ними задачи. Типичная задача параллельных корпусов включает в себя поиск по корпусу по словам или по тегам.

Для осуществления поиска по корпусу, в нем должна храниться следующая информация:

- тексты и их метаданные,
- выравнивание,
- теги и аннотации.

Теги и аннотации, наряду с выровненными текстами обычно [много ссылок] хранятся в формате XML. Но хранение информации в таком формате имеет ряд недостатков:

- избыточность (xml теги, дублирование информации)
- долгий поиск (парсинг текстового файла, да ещё и xml - с кучей лишней инфы в виде тегов)
- большой размер (можно хранить ту же самую информацию в виде бинарного файла и экономить место)
- TODO FIXME

1.2 Тексты

В данном разделе будут рассмотрены основные виды текстовых разметок, структура технических текстов, а также будет описана проблема, возникающая при автоматическом выравнивании текстов на уровне терминов.

1.2.1 Виды разметок

Морфологическая

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Синтаксическая

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Семантическая

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Метаразметка

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

1.2.2 Технические тексты и их структура

Когда кто-то называет текст «техническим» в повседневной жизни, под этим обычно понимается сложность его восприятия. В ученых кругах — наоборот, «техничность» текста означает большую доступность его обработки вследствие лишенности фигуративного языка, минимального использования переносных значений слов и возможности понимать его содержимое в буквальном смысле. [3]

Но четкого и общепринятого определения у понятия «технический текст» нет.

Тем не менее, у людей есть общее представление о том, какой текст является техническим, а какой — нет.

Исходя из предположения о том, что среднестатистический человек способен оценивать «техничность» текста, исследование [3], включало проведение массового опроса участников, в результате которого были выделены критерии, значения которых обычно выше у технических текстов.

У технических текстов преобладают следующие критерии:

- наличие заголовков,
- определение темы и фокусировка на ней,
- предоставление знаний,
- серьезность и объективность,
- логичность и последовательность,
- иерархическая организация,
- использование специализированной терминологии.

Проанализировав ряд текстов на предмет их соответствия указанным выше критериям, становится возможным выявить структуру, характерную техническим текстам, а именно:

- 1) Резюме, аннотация — краткое описание содержания текста.
- 2) Введение — знакомство с проблемой.
- 3) Теоретическая часть — более глубокое знакомство с проблемой, знакомство читателя с методологией исследования и обоснованный выбор методов, которые будут использоваться в ходе исследования.
- 4) Методология — описание применения выбранных методов в ходе исследования.
- 5) Результаты — представление полученных результатов.
- 6) Обсуждение и анализ результатов.
- 7) Заключение.

1.2.3 Проблема терминов

Выравнивание текстов на уровне секций, абзацев и предложений обычно не представляет трудности, и часто такое выравнивание можно автоматизировать. Проблемы возникают при попытке выровнять тексты на уровне терминов. Автоматически такое выравнивание произвести бывает сложно. Причина сложности автоматического выравнивания на уровне терминов будет рассмотрена ниже на примере фразеологизмов.

Набор слов в фразеологизмах может иметь разные значения в зависимости от контекста. Например, предложение «It was a piece of cake» нельзя перевести однозначно, не зная контекста, в котором оно употреблено.

В контексте

- Was it difficult?
- It was a piece of cake.

его можно перевести, как «было просто», а в контексте

- What was in the box?
- It was a piece of cake.

оно точно имеет отношение к куску пирога.

Таким образом, для корректности перевода, машинный перевод должен учитывать контекст, в котором термин употреблен. Но это и есть одна из задач, для решения которой параллельные корпуса и создаются изначально. В этом и заключается проблема терминов.

1.3 Существующие параллельные корпуса

1.4 Вывод

2 Конструкторская часть

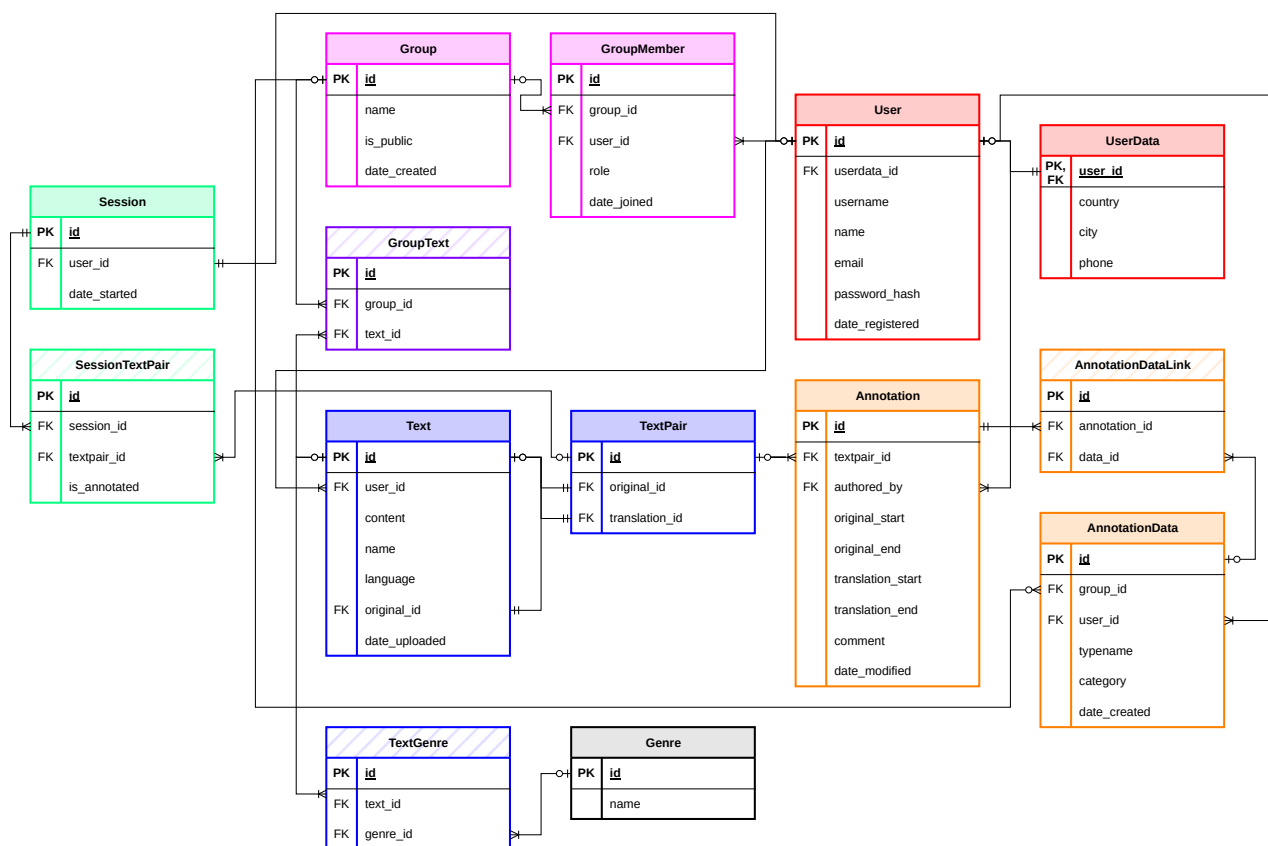


Рисунок 1 – ER-диаграмма разрабатываемой базы данных

ERD в нотации Чена: 4.

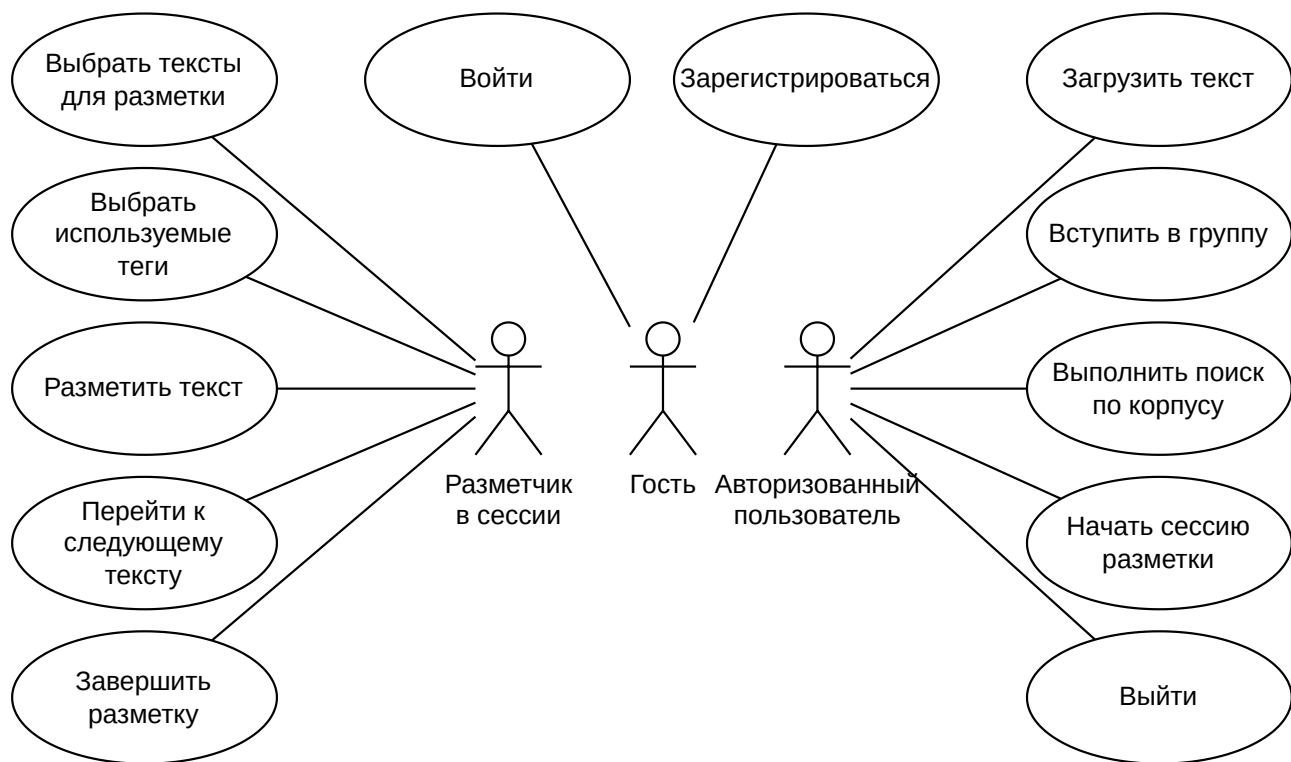


Рисунок 2 – Диаграмма сценария использования

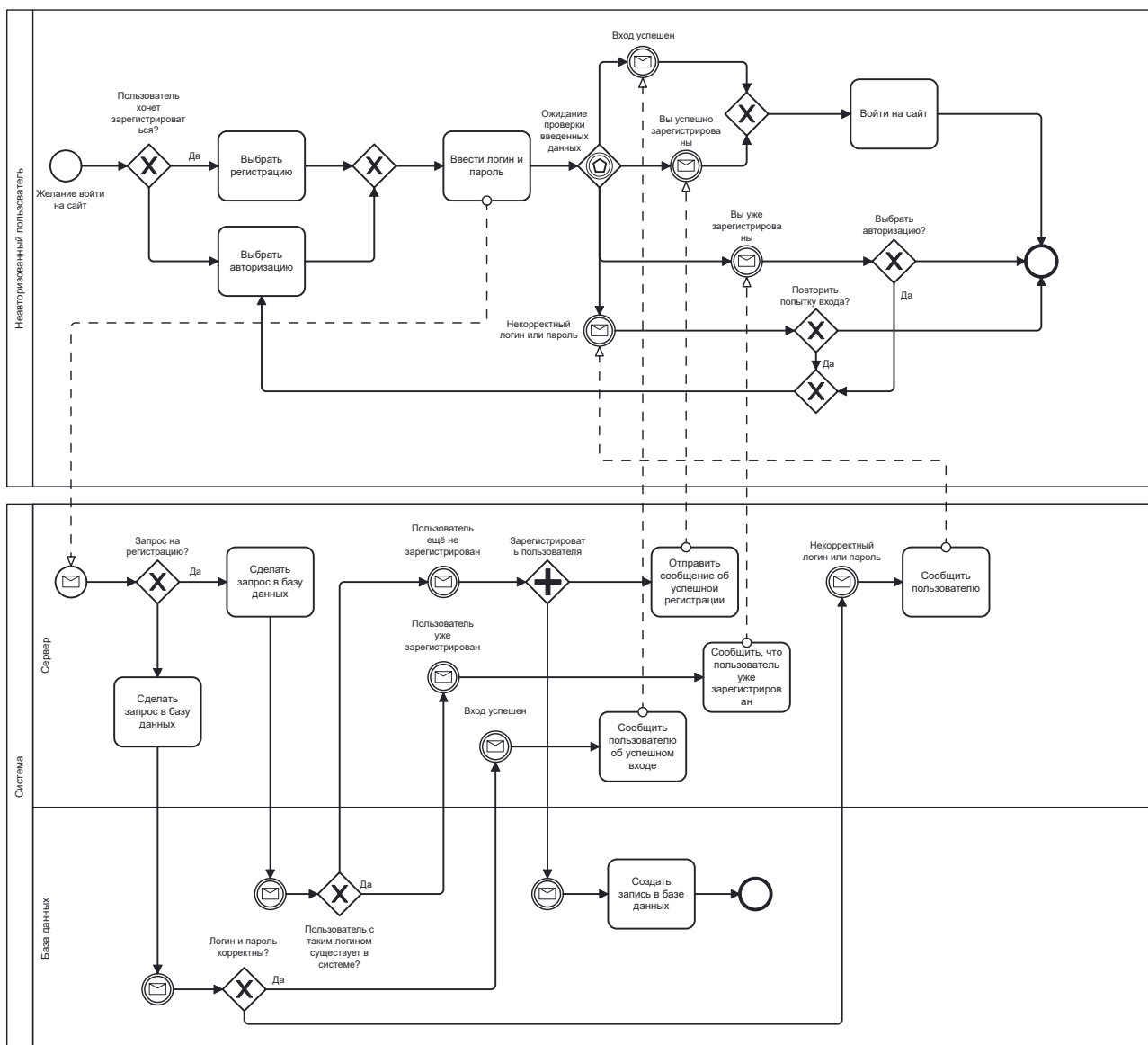


Рисунок 3 – Диаграмма процесса аутентификации пользователя

2.1 Вывод

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн., – СПб.: СПбГУ. РИО. Филологический факультет, 2013. – 148 с.
2. Параллельные корпуса текстов [Электронный ресурс]. – URL: <https://postnauka.org/video/54851> (дата обращения: 25.03.2024).
3. What is technical text? / Terry Copeck, Ken Barker, Sylvain Delisle [и др.] // Language Sciences. 1997. Т. 19, № 4. С. 391–423. – URL: <https://www.sciencedirect.com/science/article/pii/S038800019700003X>.

ПРИЛОЖЕНИЕ А

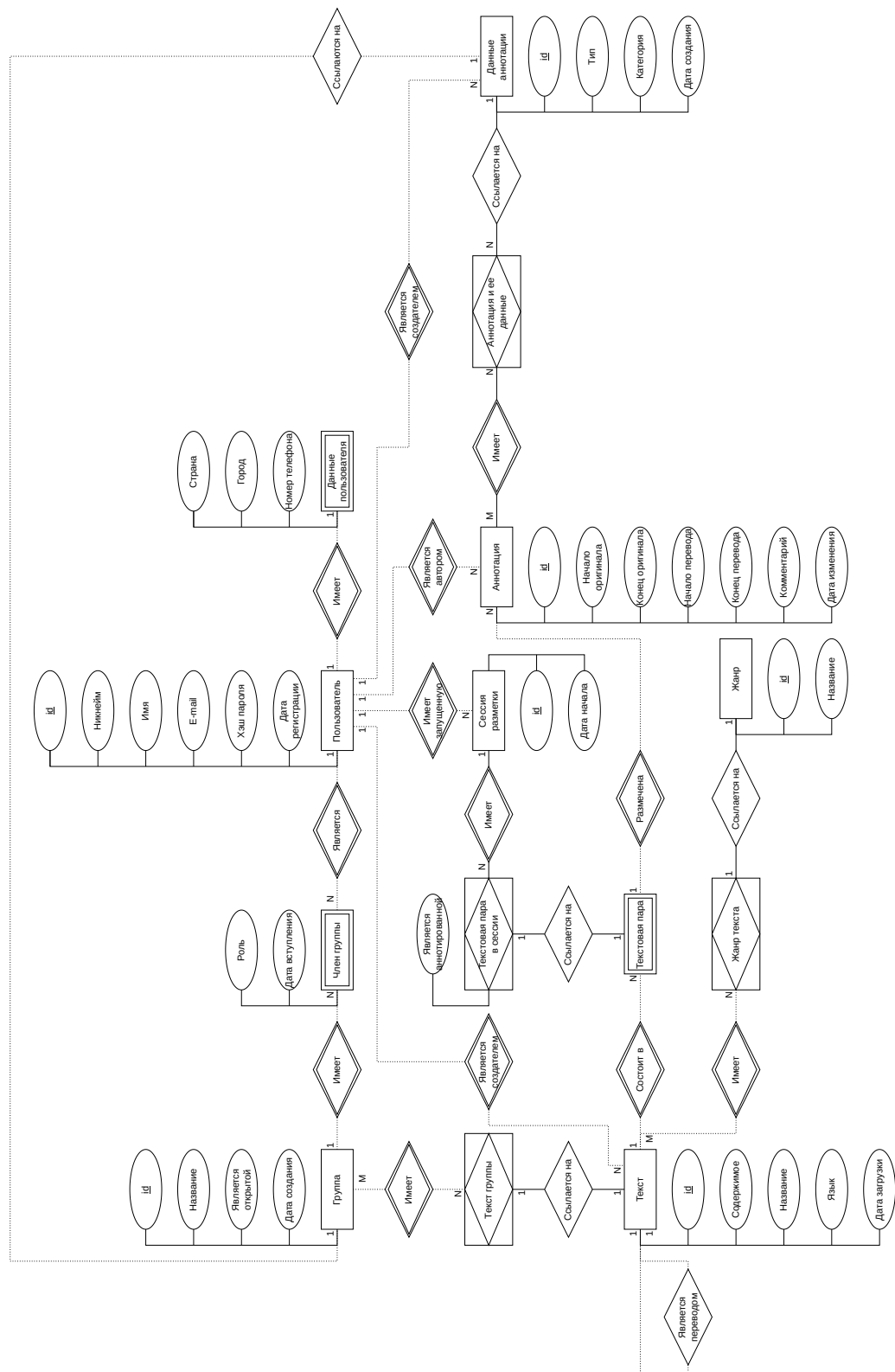


Рисунок 4 – ER-диаграмма разрабатываемой базы данных в нотации Чена