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

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

Машинный перевод (MT) представляет собой процесс перевода с одного естественного языка на другой с использованием компьютеров. Параллельные корпусы используются, как правило, для систем статистического машинного перевода, принцип работы которого заключается в анализе больших массивов параллельных текстов и в выборе для перевода наиболее часто встречающихся вариантов.

Кросс-языковой поиск информации (CLIR)

Параллельные корпуса могут использоваться для обучения моделей, способных искать информацию на одном языке и предоставлять информацию на другом.

Разработка словарей и учебных материалов

Параллельные корпуса могут использоваться при разработке словарей и учебных материалов, например, для создания примеров употребления слов

в контексте.

Лингвистические исследования

В лингвистических исследованиях параллельные корпуса могут использоваться для подсчета частот встречаемости различных языковых элементов. Статистическими методами на материале корпуса можно определить, какие слова регулярно встречаются вместе и, таким образом, могут быть отнесены к устойчивым словосочетаниям. [1, с. 81]

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 (хмл теги, дублирование информации)

— долгий поиск (парсинг текстового файла, да ещё и хмл - 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 виды текстовых разметок и структура технических текстов, а также будет описана проблема, возникающая при автоматическом выравнивании текстов на уровне терминов.

1.2.1 Виды разметок

В данном подразделе описаны некоторые виды текстовых разметок.

Морфологическая

Представление в корпусе информации о морфологических формах и значениях (часть речи, род, падеж, вид...) является самостоятельной научной проблемой. [3]

В национальном корпусе русского языка [3] морфологическая информация, приписываемая произвольной словоформе в тексте, содержательно делится на четыре типа информации:

1) Лексема (лемма), которой принадлежит словоформа (указывается «словарная запись» данной лексемы и ее принадлежность к той или иной части речи).

2) Множество грамматических признаков данной лексемы, или словоклассифицирующие характеристики (например, род для существительного, переходность для глагола).

3) Множество грамматических признаков данной словоформы, или словоизменительные характеристики (например, падеж для существительного, число для глагола).

4) Информация о нестандартности грамматической формы, орфографических искажениях и т. п.

Примеры тегов морфологической аннотации в НКРЯ [3]:

— S — существительное (яблоня, лошадь, корпус, вечность);

— A — прилагательное (коричневый, таинственный, морской);

— NUM — числительное (четыре, десять, много);

— m — мужской род (работник, стол);

— anim — одушевленность (человек, ангел, утопленник);

— sg — единственное число (яблоко, гордость);

— nom — именительный падеж (голова, сын, степь, сани, который);

— persn — личное имя (Иван, Дарья, Леопольд, Эстер, Гомер, Маугли).

Синтаксическая

Синтаксическую структуру предложения можно представить в виде дерева зависимостей, в котором каждая дуга (стрелка) идет из главного слова («хозяина») в зависимое слово («слугу») и помечена именем определенного синтаксического отношения. Каждое слово предложения, кроме одного (называемого вершиной предложения), зависит только от одного хозяина и только по одному из синтаксических отношений. Отношения связывают отдельные слова, а не словосочетания (за исключением специальных случаев). В синтаксических группах (например, в сочиненной именной группе или придаточной клаузе) один из членов группы выступает в качестве представителя группы во внешних связях, объявляется вершиной группы, а остальные члены группы синтаксически зависят от него. [3]

На рисунке ниже представлена синтаксическая разметка в двух форматах — в формате универсальных зависимостей (UD) и в формате СинТагРус.

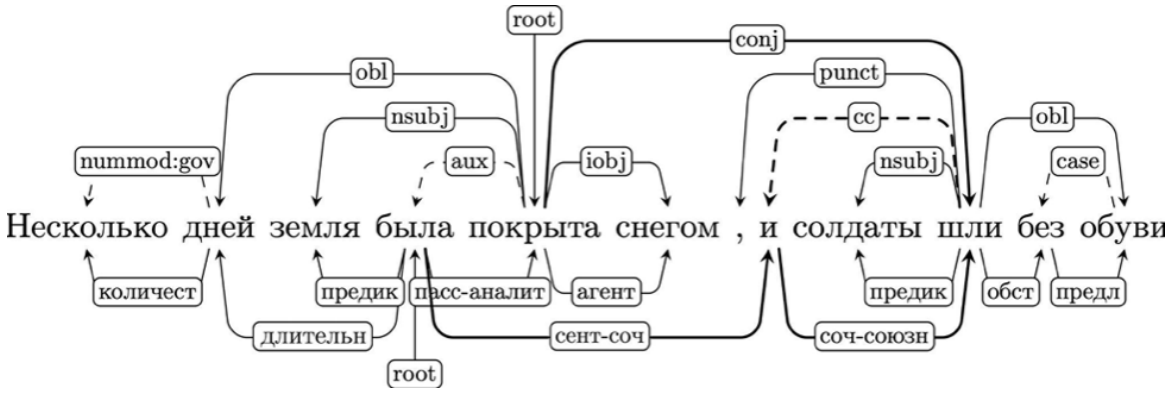

Рисунок 1 – Разметка предложения в формате UD (сверху) и СинТагРус (снизу). Стрелки отношений в сочиненной группе выделены. Функциональные отношения UD обозначены пунктирными стрелками.

Семантическая

При семантической разметке большинству слов в тексте приписывается один или несколько семантических и словообразовательных признаков, например, 'лицо', 'вещество', 'пространство', 'скорость', 'движение', 'обладание', 'свойство человека', 'диминутив', 'отглагольное имя' и т. п. В национальном корпусе русского языка [3] используется фасетная классификация, при которой одно слово может попадать в несколько классов.

Метаразметка

Под метаразметкой понимается приписывание тексту атрибутов, характеризующих обстоятельства его создания, автора, тематику, жанровые особенности и др. [3]

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 б**о**льшую доступность его обработки вследствие лишенности фигуративного языка, минимального использования переносных значений слов и возможности понимать его содержимое в буквальном смысле. [4]

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 [4], 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

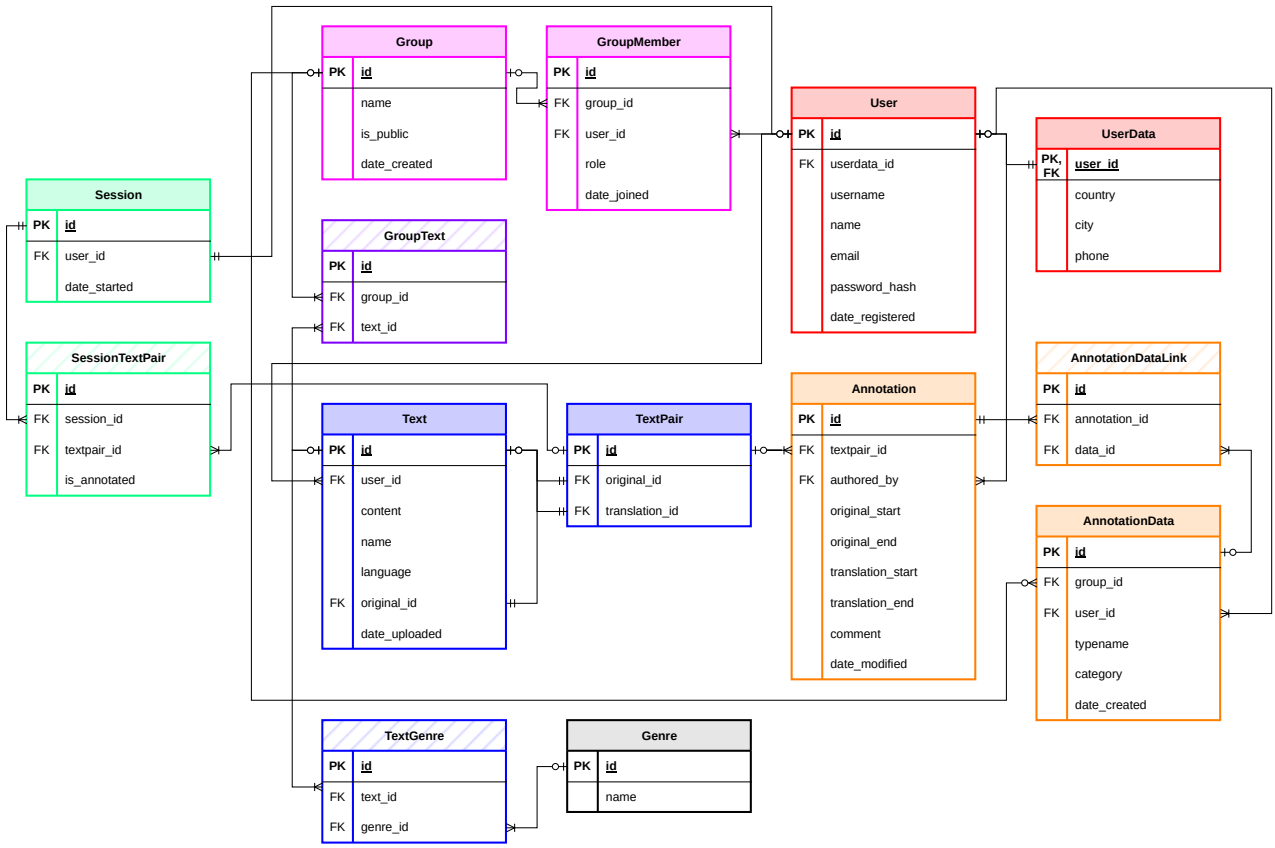

Рисунок 2 – ER-диаграмма разрабатываемой базы данных

ERD в нотации Чена: 5.

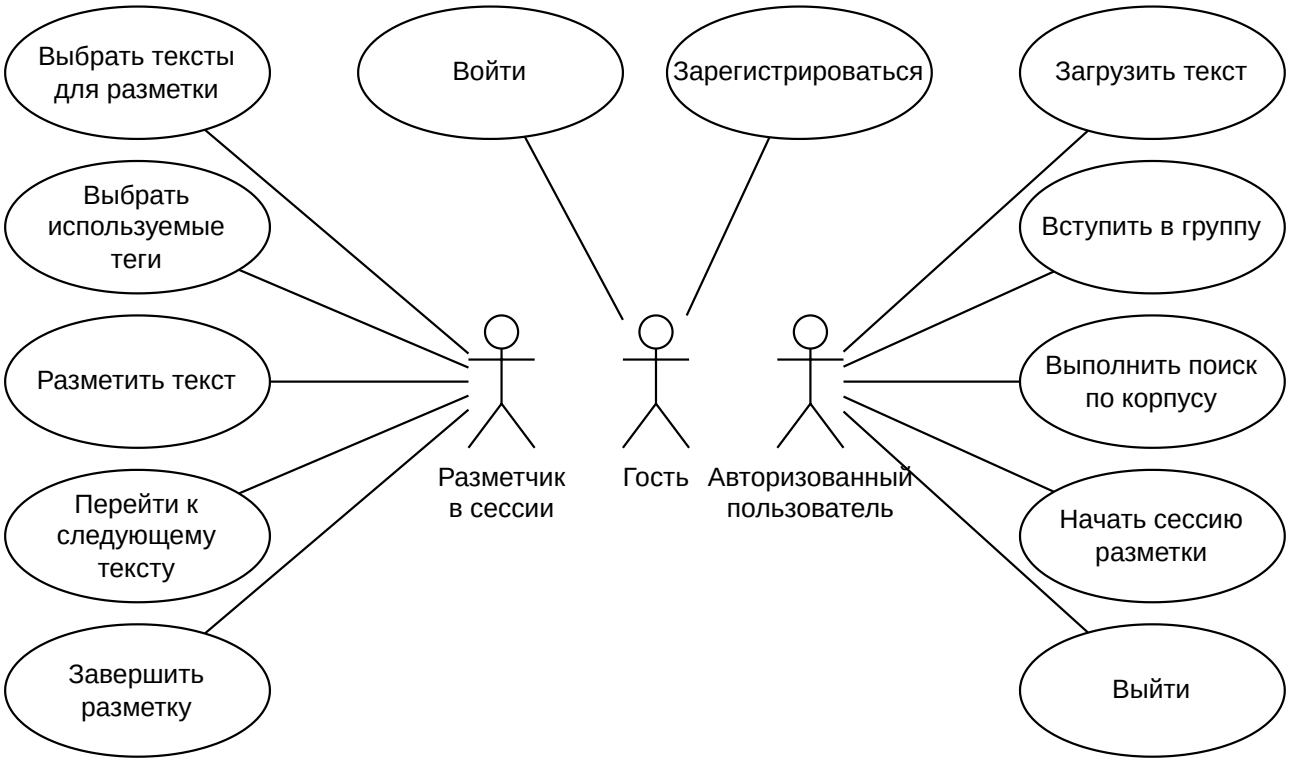

Рисунок 3 – Диаграмма сценария использования

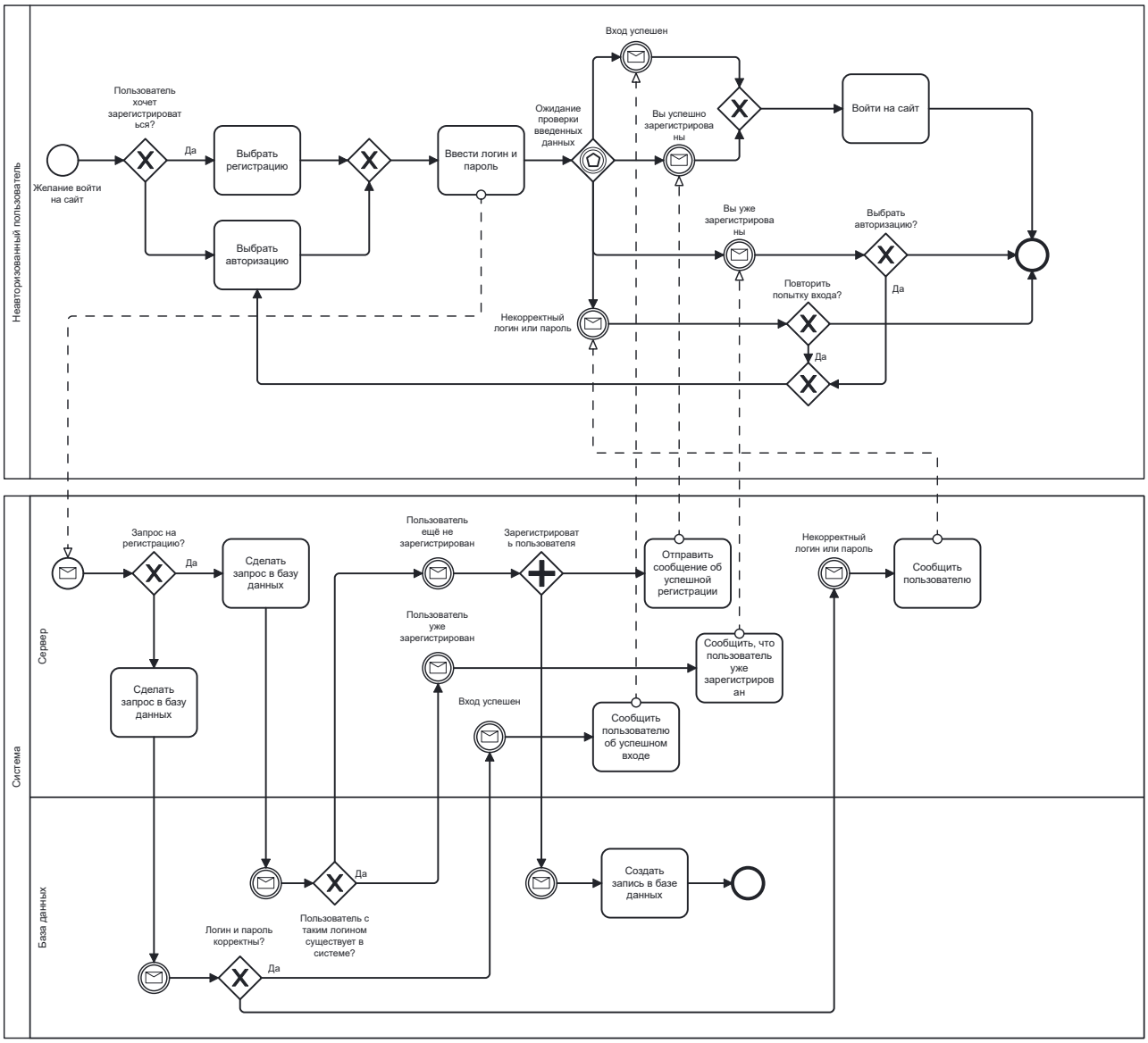

Рисунок 4 – Диаграмма процесса аутентификации пользователя

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 корпусы текстов [Электронный ресурс]. – URL: https://postnauka.org/video/54851 (дата обращения: 25.03.2024).

3. Национальный корпус русского языка [Электронный ресурс]. – URL: https://ruscorpora.ru (дата обращения: 01.03.2024).

4. 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

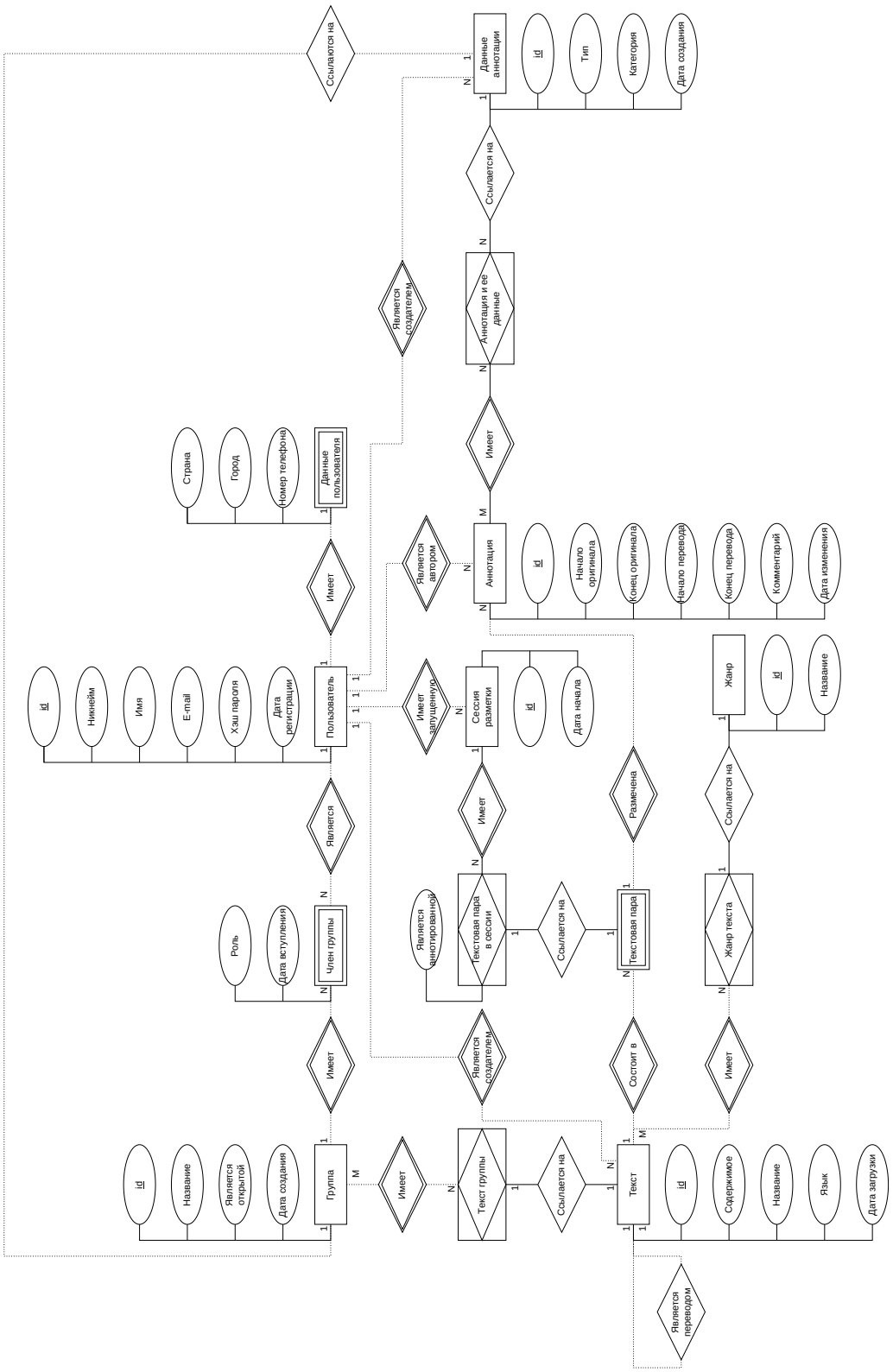

Рисунок 5 – ER-диаграмма разрабатываемой базы данных в нотации Чена