|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство образования и науки 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 Системы обработки информации и управления

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ***

***НА ТЕМУ:***

***«Модуль формирования метаграфа на основе текстового описания»***

Студент \_\_ИУ-5 84\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_А.А.Фадеев\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель ВКР **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_В.Б.Тимофеев\_\_**

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_Ю.Е.Гапанюк\_\_**

(Подпись, дата) (И.О.Фамилия)

Нормоконтролер **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_Ю.Н.Кротов\_\_\_**

(Подпись, дата) (И.О.Фамилия)

2020 г.

РЕФЕРАТ

Расчётно-пояснительная записка квалификационной работы бакалавра содержит P страниц. С приложениями объем составляет N страниц. Работа включает в себя T таблиц и I иллюстраций. В процессе выполнения было использовано S источников.

Объектом разработки является программный модуль, формирующий метаграф на основе результатов синтаксического анализа русскоязычного текста.

После синтаксического анализа исходный текст преобразуется в метаграф. Он отражает синтаксическую структуру входной последовательности и хорошо подходит для дальнейшей обработки.

В процессе выполнения квалификационной работы бакалавра были изучены различные аспекты работы с естественным языком (ЕЯ) и её автоматизации.

Расчётно-пояснительная записка состоит из двух частей.

Постановка задач разработки рассматривает данную предметную область, процессы, подлежащие автоматизации при создании синтаксического анализатора и метаграфа, различные подходы и алгоритмы, которые для этого используются.

Конструкторско-технологическая часть включает выбор языка программирования для разработки программного изделия, проектирование, правил, базы данных, алгоритмов синтаксического анализа, формирования метаграфа и интерфейса пользователя.

В расчётно-пояснительную записку включены 3 приложения.

СОДЕРЖАНИЕ

[РЕФЕРАТ 2](#_Toc42547834)

[СОДЕРЖАНИЕ 3](#_Toc42547835)

[ВВЕДЕНИЕ 5](#_Toc42547836)

[1. ПОСТАНОВКА ЗАДАЧ РАЗРАБОТКИ 6](#_Toc42547837)

[1.1. Общетехническое обоснование разработки (Ваш вариант) 6](#_Toc42547838)

[1.1.1. Постановка цели работы 6](#_Toc42547839)

[1.1.2. Постановка задач 6](#_Toc42547840)

[1.1.3. Описание предметной области 7](#_Toc42547841)

[1.1.4. Перечень процессов, подлежащих автоматизации 9](#_Toc42547842)

[1.1.5. Выбор критериев качества 10](#_Toc42547843)

[1.1.6. Анализ аналогов и прототипов 12](#_Toc42547844)

[1.2. Анализ существующих алгоритмов и внедрение отличительных особенностей 14](#_Toc42547845)

[2. КОНСТРУКТОРСКО-ТЕХНОЛОГИЧЕСКАЯ ЧАСТЬ 21](#_Toc42547846)

[2.1. Выбор программных средств 21](#_Toc42547847)

[2.2. Выбор аппаратных средств 22](#_Toc42547848)

[2.3. Выбор языка разработки 23](#_Toc42547849)

[2.4. Структура программного изделия 24](#_Toc42547850)

[2.5. Проектирование структуры правил 26](#_Toc42547851)

[2.6. Проектирование модели базы данных 30](#_Toc42547852)

[2.7. Разработка программного продукта 46](#_Toc42547853)

[2.7.1. Разработка интерфейса пользователя 46](#_Toc42547854)

[2.7.2. Разработка алгоритмов синтаксического анализа текста 47](#_Toc42547855)

[2.7.3. Разработка структуры классов модуля построения метаграфа 48](#_Toc42547856)

[2.7.4. Разработка алгоритма преобразования результатов синтаксического анализа в метаграф 50](#_Toc42547857)

[2.8. Результаты опытной эксплуатации 52](#_Toc42547858)

[ЗАКЛЮЧЕНИЕ 53](#_Toc42547859)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 54](#_Toc42547860)

[ПРИЛОЖЕНИЕ А ГРАФИЧЕСКИЕ МАТЕРИАЛЫ 55](#_Toc42547861)

[ПРИЛОЖЕНИЕ В ТЕХНИЧЕСКОЕ ЗАДАНИЕ 58](#_Toc42547862)

[ПРИЛОЖЕНИЕ В ПРОГРАММА И МЕТОДИКА ИСПЫТАНИЙ 64](#_Toc42547863)

# ВВЕДЕНИЕ

Квалификационная работа на тему «Формирование метаграфа на основе текстового описания» посвящена разработке модуля для построения метаграфа по результатам синтаксического анализа исходного русскоязычного текста.

В данной области множество научных работ и различных продуктов, дошедших до простого пользователя: электронные переводчики, распознаватели текста, поисковые системы, вопросно-ответные системы. Но почти все доступные сторонним пользователям системы не понимают слова так, как это делает человек, а лишь обрабатывают цепочки символов, используя статистику.

При работе с ЕЯ основными задачами являются разбор и анализ входного текста или синтез выходных предложений. Анализа текстов на ЕЯ состоит из 4-х этапов: морфологический, синтаксический, семантический и прагматический.

В данной квалификационной работе подробно рассматривается синтаксический этап разбора входного текста. Морфологический этап разбора проводится за счёт сторонних программных средств, о которых говорится в пункте 2.1 данной работы.

Морфологический анализ включает определение морфологических характеристик слова, таких как часть речи, род, число, падеж и т.д.

Синтаксический анализ, который содержит определение синтаксических ролей слов и словосочетаний в предложении. Результатом анализа является метаграф синтаксических зависимостей слов, который отражает структуру их взаимодействия в рамках одного предложения.

1. ПОСТАНОВКА ЗАДАЧ РАЗРАБОТКИ
   1. Общетехническое обоснование разработки
      1. Постановка цели работы

В настоящее время существует множество систем, каким-либо образом анализирующих содержание текста. Однако почти все открытые для пользователей системы не анализируют текст как таковой, а лишь исследуют «важность» тех или иных слов внутри него, проводя максимально примитивный семантический анализ. Синтаксический анализ предваряет семантический, поэтому его роль в анализе текстов на ЕЯ не менее важна.

В связи с этим цель работы формулируется следующим образом: предоставление возможности синтаксического анализа текста на русском языке и построения структуры, которая отражает синтаксическую структуру входной последовательности, без сколь угодно простого семантического анализа. А также предоставлении возможности консольной визуализации метаграфа-результата синтаксического разбора предложений текста.

* + 1. Постановка задач

В соответствии с целью данной работы, подразумевающей под собой создание синтаксического анализатора без каких-либо функций семантического анализатора, задачами проектирования становятся:

* Создание алгоритма фрагментации сложных предложений;
* Создание алгоритма поиска грамматической основы предложения;
* Создание структуры правил выделения синтаксических конструкций;
* Создание алгоритма применения правил для построения синтаксических групп (СГ) на фрагментах простых предложений в составе сложных;
* Формирование структуры синтаксического метаграфа;
* Создание алгоритма построения метаграфа;
* Разработка модуля консольной визуализации метаграфа.
  + 1. Описание предметной области

Текст – довольно однородная структура, а слова, являющиеся его элементами, более-менее равноправны, поскольку в каком-либо тексте мы, люди без помощи компьютера, не можем придать одному слову значение всего текста.

В настоящее время количество и, следовательно, объём электронных документов увеличивается с гигантской скоростью. В связи с этим каждый день возрастает потребность в обработке неструктурированной текстовой информации, повышению качества и эффективности имеющихся методов и разработке универсальных технологий обработки текста. В настоящее время актуально фактически любое направление, связанное с облегчением этой глобальной задачи: поиск информации, фильтрация, кластеризация документов, поиск ответов на вопросы, автоматическое аннотирование документа и группы документов, поиск похожих документов и дубликатов, сегментирование документов, формирование баз знаний на их основе и многое другое.

Сейчас такая обработка ведётся в основном статистическими методами, подкреплёнными, например, технологией нейронных сетей. Уже несколько десятков лет ведётся научная работа по поиску оптимальных алгоритмов семантической обработки текста (семантических анализаторов), которая позволит компьютеру понять текст.

Если обобщить различные определения семантического анализатора, то данное понятие становится очень простым. Семантический анализатор – это программное средство, которое, благодаря своей алгоритмической мощи в общем случае позволяет понимать, анализировать, находить истину и ложь, недосказанности и противоречия, аннотировать и редактировать тексты, находить в них смысл и знания, организовать диалог с пользователями или отвечать на их вопросы по каким-либо текстам и т.п. Однако первоочередной задачей семантических анализаторов всегда есть и будет поиск знаний в различных текстовых источниках информации, представленных в форме естественного языка.

Чтобы выполнить эту задачу, необходимо произвести синтаксическую обработку текста. Существует несколько подходов к реализации синтаксических анализаторов. Выбор использованного метода подробно описан в пункте 1.3 данной работы. Метод предполагает явное разделение процесса анализа текста на следующие модули: морфологический анализ, синтаксический анализ, семантический анализ и прагматический анализ.

В данной работе рассматривается процесс синтаксического анализа с использованием готовых решений морфологического анализа.

В процессе исследования и разработки были выделены следующие понятия, необходимые для реализации выбранного метода синтаксического анализа:

* Синтаксический анализатор (СА) – программа, которая производит синтаксический анализ текста без его сколь угодно простого семантического анализа;
* Граммема – компонент грамматической категории, представляющий собой по своему значению видовое понятие по отношению к значению грамматической категории как понятие родовому;
* Часть предложения – простое предложение с одной грамматической основой в составе сложного, исходного предложения текста;
* Синтаксическая группа (СГ) – группа синтаксически связанных слов, среди которых выделено главное, которое определяет роль группы;
* Входная СГ – СГ, рассматриваемая как один двух из входов правила;
* Правило – принцип обработки двух входных синтаксических групп, условий согласования и преобразования двух СГ в результирующую СГ;
* Тип входной СГ – константа, которая определяет синтаксическую роль (подчиненная или главная) двух входящих в правило СГ во время его применения.
* Вид СГ – характеристика каждой СГ, которая определяет набор правил, применимых к данной СГ.
  + 1. Перечень процессов, подлежащих автоматизации

При разработке модуля морфологического анализа текста следующие процессы подлежат автоматизации:

* Фрагментация текста на абзацы, абзацев – на предложения, а предложений – на лексемы (слова и знаки препинания);
* Формирование словаря, сохраняющего информацию о строении текста и морфологическую информацию о словах, то есть информацию о вариантах морфологического разбора каждой лексемы предложения.

При разработке модуля синтаксического анализа текста следующие процессы подлежат автоматизации:

* Формирование списка союзов каждого предложения;
* Фрагментация сложного предложения на простые с учётом информации о союзах и знаках препинания;
* Поиск грамматической основы каждой части предложения;
* Поиск вариантов цепочек словоформ следуя результатам морфологического разбора для лексем каждой части предложения;
* Применение правил к очередному варианту части предложения;
* Запоминание успешных попыток анализа части предложения;

При разработке модуля формирования метаграфа, как результата синтаксического разбора предложения:

* Формирование списка метаграфов на основе информации об абзацах, и положении предложений текста в абзацах, а абзацев в тексте.
  + 1. **Выбор критериев качества**

Проблема понимания компьютером естественного языка широко распространена, многие прикладные системы используют в процессе работы синтаксические анализаторы, принципы работы которых разобраны более-менее хорошо. Этим обусловлено наличие аналогов. Для сравнения выбраны системы Solarix и AOT. Варианты синтаксических анализаторов представлены в таблице 1.

**Таблица 1. Варианты синтаксических анализаторов**

| **№ п/п** | **Название варианта** | **Обозначение** |
| --- | --- | --- |
| 1. | Solarix | В1 |
| 2. | AOT | В2 |
| 4. | Разрабатываемый модуль | В3 |

Так как мы уже выбрали алгоритмы разрабатываемого СА, мы можем сравнить его с другими, готовыми СА. Для разрабатываемого программного изделия приоритетными являются критерии качества, приведенные в таблице 2.

**Таблица 2. Критерии сравнения аналогов и прототипов**

| **Критерий** | **Описание** | **Обозначение** |
| --- | --- | --- |
| Наличие графического интерфейса | Критерий показывает, насколько пользователю будет удобно рассматривать и анализировать результаты анализа по картинке на экране | К1 |
| Стоимость | Стоимость программного модуля для его использования в собственной программе | К2 |
| Платформа | Платформа для которой предназначена система | К3 |
| Сложность эксплуатации | Сложность использования модуля в рамках собственной программы | К4 |
| Форматы выходных данных | Качество формата данных, генерируемых системой | К5 |
| Открытый исходный код | Доступность исходного кода программы, предоставляющая возможность собственной модификации | К6 |

Показатели различных вариантов по данным критериям представлены в таблице 3.

**Таблица 3. Показатели различных вариантов по данным критериям**

| **Варианты** | **К1** | **К2** | **К3** | **К4** | **К5** | **К6** |
| --- | --- | --- | --- | --- | --- | --- |
| В1 | 3 | 1 | 5 | 4 | 3 | 4 |
| В2 | 2 | 1 | 3 | 4 | 3 | 5 |
| В3 | 4 | 5 | 5 | 3 | 5 | 5 |

* + 1. Анализ аналогов и прототипов

Для анализа выбран метод базового критерия. В связи с этим методом обозначенным выше критериям назначены весовые коэффициенты. Результаты приведены ниже в таблице 4.

**Таблица 4. Проранжированные критерии качества**

| **№**  **п/п** | **Название критерия** | **Весовой коэффициент** |
| --- | --- | --- |
| 1. | Наличие графического интерфейса | 2 α |
| 2. | Стоимость | 5 α |
| 3. | Платформа | 6 α |
| 4. | Сложность эксплуатации | 3 α |
| 5. | Форматы выходных данных | 4 α |
| 6. | Открытый исходный код | 5 α |

Определим величину α, рассчитав значения весовых коэффициентов по формуле 1 [8].

, (1)

где *п* – общее количество весовых коэффициентов; *ai*  – весовые коэффициенты.

Подставив значения в формулу 1 получаем:

В таблице 5 отображены результата расчёта весов методом базового критерия.

**Таблица 5. Расчет весов методом базового критерия**

| **Критерий** | **Коэффициент важности (ki)** | **Вес (αi)** |
| --- | --- | --- |
| К1 | 2 | 0.08 |
| К2 | 5 | 0,2 |
| К3 | 6 | 0,24 |
| К4 | 3 | 0,12 |
| К5 | 4 | 0,16 |
| К6 | 5 | 0,2 |

Для вычисления локальных критериев необходимо произвести нормировку значений критериев.

Нормированные значения приведены в таблице 6.

**Таблица 6. Нормированные значения локальных критериев**

| **Критерий** | **Вес (αi)** | **В1** | **В2** | **В3** |
| --- | --- | --- | --- | --- |
| К1 | 0,08 | 0,75 | 0,5 | 1 |
| К2 | 0,2 | 0,2 | 0,2 | 1 |
| К3 | 0,24 | 1 | 0,6 | 1 |
| К4 | 0,12 | 1 | 1 | 0,75 |
| К5 | 0,16 | 0,6 | 0,6 | 1 |
| К6 | 0,2 | 0,8 | 1 | 1 |
| Итог | | 0,71225 | 0,64 | 0,97 |

При этом итоговые значения рассчитаны по формуле 2.

, (2)

Лучший вариант выберем по формуле 3.

, (3)

где *m* – общее количество вариантов.

Таким образом получили, что разрабатываемый модуль имеет предпочтение над имеющимися аналогами. Расчет показывает, что есть основания для разработки данного программного модуля.

* 1. Анализ существующих алгоритмов и внедрение отличительных особенностей

Для построения архитектуры синтаксического анализатора, которая отвечает всем требованиям, указанным в индивидуальном задании, необходимо проанализировать все решения, которые используются в разработке синтаксических анализаторов.

Детальный анализ приведён в трудах [1, 2]. В них говорится о различных стратегиях при анализе синтаксических конструкций с разной степенью детализации.

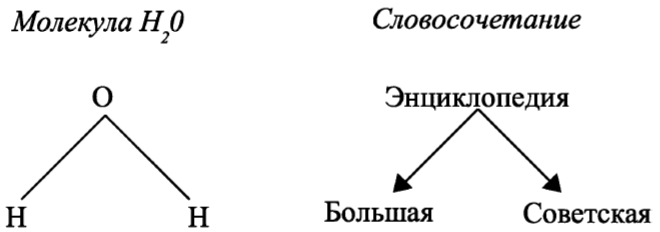
Наиболее четкие границы есть между тремя подходами к описанию синтаксиса:

1) Грамматики зависимостей;

2) Грамматики непосредственных составляющих;

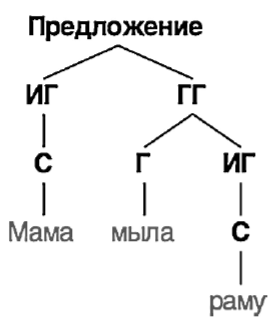
3) Комбинированные теории, например, теория синтаксических групп.

Основной принцип подхода грамматики зависимостей заключается в том, что каждое слово предложения составляет иерархию. Каждое слово является своего рода атомом в молекуле. Это означает что само предложение в результате анализа представляется в виде дерева зависимостей или, другими словами, графа зависимостей. Таким образом синтаксические связи устанавливают между словами отношения зависимости. Одно из двух слов является главным, а другое зависимым. Например, в словосочетании «Большая Советская Энциклопедия» имеется две связи, изображённые на рисунке ниже (см. рис.1). Эти связи образуют отношения зависимости: в обеих связях главной является словоформа энциклопедия. Словоформы «Большая» и «Советская» оказываются зависимыми.

Рис. 1. Строение взаимосвязей в «грамматике зависимостей»

Как мы видим, узким местом этого подхода является невозможность построения какой-либо другой связи, кроме подчинительной.

Непосредственные составляющие (НС, фразы) – на рисунке ниже (см. рис.2) приведён пример структуры предложения "Мама мыла раму". У этого предложения имеются НС: "мама" и "мыла раму". В свою очередь, "мыла раму" имеет составляющие "мыла" и "раму", то есть эти НС не входят непосредственно в предложение, а являются компонентами его более-менее крупной части. По этой схеме ни одна словоформа не будет являться НС целого предложения, но будет являться НС по отношению к тем или иным его составляющим.

  
Рис. 2. Грамматика непосредственных составляющих.

По данной схеме предложение состоит из одной или нескольких НС. Каким образом данные НС-грамматика оперирует такими понятиями, как классы НС, например, предложения, именная группа или глагольное группа. Однако названия этих классов никак не обосновываются частями речи тех словоформ, которые входят в его состав.

Именно в подходе непосредственных составляющих появляется такое понятие как восходящие алгоритмы или нисходящие алгоритмы. Подробное описание восходящих и нисходящих алгоритмов описано в работе [2]. Для данной работы «восходящий» способ был выбран именно потому, что он отличается относительной простотой в плане построения правил нахождения взаимосвязи между различными НС. Основной принцип такого вида алгоритма заключается в том, что в предложении ищется наименее значимая с синтаксической точки зрения словоформа, и она соединяется с более значимой по специальным правилам, например, "частица+глагол" или "предлог+существительное". Таким образом мы получаем возможность отсечения неверных вариантов на том этапе разбора предложения, когда мы не обработали все словоформы и не порождаем множество заведомо неверных вариантов. Таким образом мы можем экономить время работы и ресурсы компьютера. Одним из основных минусов такого вида алгоритмов является порождение так называемых эллипсисов или рекурсивных грамматик. Эллипсис – возникновение пропусков некоторых слов в предложении. В таком случае мы должны восстанавливать потерянные слова из контекста. Примеры предложений с эллипсисом: "Он нравится Маше, а она – ему".

Комбинированные алгоритмы нивелируют минусы друг друга за счёт того, что в них может реализовываться разный подход к использованию восходящего алгоритма и нисходящего алгоритма. Например, сначала может быть разобранной восходящем алгоритмом именная группа подлежащего в предложении, а нисходящем – группа сказуемого. За счет использования принципов обоих подходов комбинированные алгоритмы находятся по середине между восходящими и нисходящими алгоритмами в плане производительности и скорости. Однако они впереди нисходящих и восходящих алгоритмов в плане точности разбора предложений, поскольку они могут реализовывать разбор эллипсисов, рекурсивных грамматик или других сложных случаев предложений, которые недоступны для разбора обычным восходящим или нисходящим алгоритмам.

На рисунке 3 изображены общие схемы работы трех описанных выше алгоритмов. Число рядом с НС показывает порядок нахождения зависимостей.

  
Рисунок 3. Порядок работы нисходящих, восходящих и комбинированных алгоритмов

Еще одним минусом восходящих и нисходящих алгоритмов НС является их неспособность анализировать те НС, которые разделены другими НС. Для устранения недостатка подобных алгоритмов А. В. Гладким была разработана теория синтаксических групп (ТСГ), которая реализуется в разрабатываемой системе.

Этот подход является представителем комбинированных теорий анализа предложения [2]. Он предполагает использование восходящего алгоритма, но уже применительно не к непосредственным составляющим, а к синтаксическим группам (СГ). ТСГ допускает включение в структуру предложения тех групп, которые участвуют в отношении зависимости «целиком», а не посредством одной словоформы. Таким образом этот подход позволяет легко обрабатывать разрывные составляющие.

Независимо от выбранного подхода главная проблема синтаксического анализа связана с неоднозначностью языковых единиц. Помимо очевидной многозначности слов (лексической неоднозначности), есть и менее очевидная многозначность грамматических форм (морфологическая неоднозначность): ср., например, например, числи тельное три и форму повелительного наклонения единственного числа глагола тереть, или форму именительного падежа слова стекло и форму прошедшего времени, единственного числа, среднего рода глагола стекать. Морфологическая неоднозначность может показаться редким явлением, но только потому, что она редко бывает очевидной.

Больше всего проблем вызывает, однако, не морфологическая, а синтаксическая неоднозначность. Из-за морфологической неоднозначности одна и та же фраза может иметь несколько разных пониманий, которые соответствуют разным структурам. Более того, одна и та же фраза может быть понята по-разному даже при однозначности всех её словоформ («Вася встретил Петю в коридоре» и «Вася встретил Петю в костюме»). В действительности, обе фразы имеют, по крайней мере, три возможные синтаксических структуры. Им соответствует три разные ситуации. Для первой фразы они таковы:

* Вася встретил Петю, и это произошло в коридоре;
* Вася встретил Петю, одетого в коридор;
* Вася, одетый в коридор, встретил Петю.

Чтобы найти верный вариант разбора целого предложения, можно попытаться перебрать все комбинации вариантов разбора его частей. Проблема в том, что при компьютерном анализе это приводит к так называемому комбинаторному взрыву.

Как уже видно из примера даже у самого короткого предложения может быть несколько интерпретаций на уровне синтаксиса и лексики. Поэтому, если стоит цель выдать единственно верный вариант синтаксического разбора, необходимо решить задачу, по сложности сопоставимую с задачей коммивояжера. Однако поиск верного решения среди нескольких вариантов –задача семантического характера, поэтому ее решение идет в разрез с целью данного проекта.

Выбранный алгоритм СА предполагает, что каждая СГ части предложения равноправна любой другой. Исключения составляют слова грамматической основы. Если слово – часть грамматической основы, то оно не может быть подчинено, а слова, подчиняемые словам грамматической основы, присоединяются с ней на предпоследнем шаге. Таким образом любое слово, кроме грамматической основы в теории может быть связано с любым другим словом предложения. Само предложение считается проанализированным, когда каждое слово, кроме грамматической основы стало подчиненным. Тогда грамматическая основа связывается в предложение одним из трёх уникальных правил: только «Подлежащее», только «Сказуемое» и «Подлежащее+Сказуемое или Сказуемое+Подлежащее».

Подробное описание правил представлено в пункте 2.5 работы. Подробное описание структуры программного изделия и конкретных алгоритмов представлено в пункте 2.4 и 2.7 работы.

1. КОНСТРУКТОРСКО-ТЕХНОЛОГИЧЕСКАЯ ЧАСТЬ
   1. Выбор программных средств

Модуль кроссплатформенный, поэтому для использующих его ПК нет ограничений по предустановленным на ним ОС.

Программа состоит из нескольких функций, работающих над разными задачами, порядок выполнения которых определяется конкретным подходом, оба которых описаны выше.

Для использования любого из этих двух принципов нужно понимать, какие задачи ставит перед собой проект. Но независимо от этих задач предварительным этапом работы любого синтаксического анализатора (СА) является токенизация текста на предложения, предложений – на лексемы: слова и знаки препинания. Для этого необходимо использовать наиболее точные библиотечные программные средства, написанные на языке Python 3:

* Разделение на предложения выполняются с помощью библиотеки ru\_sent\_tokenize. Она предназначена и разработана для разделения русского текста на предложения с учётом специфики русского языка.
* Для разделения предложений на слова используется библиотека NLTK для русского языка с предварительной загрузкой соответствующих модулей и словарей русского языка.
* Для реализации морфологического разбора лексем предложения используется библиотека pymorphy2. Она написана для морфологического анализа русского языка с использованием открытых материалов корпуса словарей русского языка OpenCorpora, обогащением которого может заняться любой желающий.

NLTK – библиотека, удачно подходящая для разбора английского языка. Предложения русского языка она разделяет с точностью 95,5 процента. ru\_sent\_tokenize справляется с этой же задачей в 99 и более процентах случаев, а также скорость выполнения задачи более, чем в полтора раза выше. Именно поэтому ru\_sent\_tokenize был выбран для разделения текста на предложения. NLTK, настроенный на морфологический анализ русского языка отлично справляется с токенизацией по словам.

В данном проекте база данных отвечает исключительно за хранение данных о правилах, и необходима лишь для того, чтобы хранить информацию о разобранных словах предложения. То есть каждый успешный результат анализа предложения записывается в Базу Данных.

Выбрана СУБД MySQL. Она оптимизирована под использование небольших приложений. В будущем с расширением возможностей СА не возрастет нагрузка на БД, нет нужды выбирать более быстрые, надежные базы данных, поскольку в ней не хранятся данные приложения или пользователей.

* 1. Выбор аппаратных средств

Модуль планируется для использования в других системах в качестве компонента.

Минимальные необходимые характеристики:

* процессор Intel Core i3 или более мощный (или совместимый с ним);
* рекомендуемый объем оперативной памяти 2048 мегабайт (МБ);
* минимальный объем 1024 МБ;
* 100 MB свободного места на жестком диске для системы и от 500 МБ до 1 ГБ для приложения, базы данных и демонстрационных файлов;
* монитор;
* клавиатура;
* мышь или совместимое указывающее устройство.
  1. Выбор языка разработки

Алгоритмическим задачам, стоящим перед проектом, удовлетворяет язык программирования Python 3. Выбран он был по следующим причинам:

* Этот язык прост;
* Реализация алгоритмов не требует множества строк кода, то есть лишнего времени на написание тонких моментов проекта;
* Онлегок и прост в использовании, интеграции в многомодульный проект;
* Язык ориентирован на решение задач общего направления, то есть не ограничивает возможности реализации СА;
* Широкая распространенность и большое количество стандартных и сторонних библиотек, которые могут оптимально упростить и ускорить разработку;
* Все перечисленные в пункте 2.1 средства используют именно его;
* Структура хранения данных идеально походит под временное хранение информации об анализе текста в течение работы программы.

Этот ЯП, однако, сложно использовать для параллельных вычислений. Куда лучше для анализа нескольких предложений сразу подходит, например, функциональный ЯП Golang, однако используемые сторонние средства не поддерживают этот ЯП. В случае его использования как основного языка разработки потребуется написать дополнительный модуль, который обеспечит взаимодействие кода на Python 3 и на Golang.

* 1. Структура программного изделия

Данный программный продукт состоит из трех логически разделенных модулей:

* Морфологический анализатор;
* Синтаксический анализатор;
* Формирование метаграфа.

Морфологический анализатор включает в себя фрагментацию текста, основная задача которого состоит в выделении абзацев в составе текста, предложений в составе абзацев, слов и знаков (лексем) препинания в составе предложения, а также в морфологическом анализе полученных лексем. На выходе данного модуля получаются фрагменты, которые готовы для синтаксического анализа.

В связи с выбранной целью было выбран подход, игнорирующий так называемый «комбинаторный взрыв». Данная проблема стоит перед всеми синтаксическими анализаторами.

Каждое слово имеет несколько допустимых правилами форм, называемых словоформами. Каждая словоформа имеет одно и несколько вариантов набора морфологических характеристик. Например, слово «её» имеет само по себе более 20 вариантов разбора. Поэтому разбор любого предложения, даже абсолютно правильный с точки зрения правил русского языка, влечёт за собой размножение вариантов разбора предложения (предложение из 7 слов может быть разобрано, например, 10-ю и более разными способами).

Чтобы решить эту проблему существующие системы пытаются использовать внедрение элементов семантического анализатора на синтаксический. Это влечёт за собой многократное возрастание сложности и временной дороговизны создания анализатора, нивелирование разницы между этими двумя этапами разбора. В данном проекте используется другая схема: морфология, синтаксис, семантика и прагматика текста полностью разделяется.

Это позволяет не только абстрагировать конкретный уровень от проблем другого, но и сосредоточиться на тонкостях проблем конкретного уровня. Например, все мы знаем, что предлоги не имеют морфологических характеристик, которые позволяют использовать их в сочетании с другими словами. Однако, мы все знаем, что, например, предлог «по» употребляется только со словами в дательном падеже. Таким образом, «обработка предлогов» позволит сократить количество вариантов морфологического разбора слов в 2-3 раза, а для неизменяемых слов (например, «пальто») ровно в 6 раз. Этот этап может предшествовать синтаксическому этапу, этапу поиска союзов предложения и грамматической основы.

После получения морфологической информации (с помощью библиотек ru\_sent\_tokenize, NLTK и pymorphy2) о составе предложения начинается этап синтаксического анализа, который проходит в два этапа (подробное описание в пункте 2.7.2):

1. Производится разделение сложных предложений на простые;

2) Анализ каждой части предложения с учётом всех возможных вариантов грамматической основы данной части, с учётом приоритета правил и т.д.;

Результат анализа предложения сохраняется в объекте метаграфа, исходя из следующих представлений (подробное описание в пункте 2.7.3):

* 1 предложение – один метаграф;
* 1 часть – одна метавершина;
* 1 слово – одна вершина;
* 1 связь простых предложений в составе сложного – одно ребро метаграфа;
* 1 связь между 2-мя словами – одно ребро метавершины.
  1. Проектирование структуры правил

В связи с выбранным методом синтаксического анализа, правила должны определять взаимодействие двух синтаксических групп. Так как алгоритм анализа можно охарактеризовать как восходящий, каждое правило имеет приоритет применения. Правила хранятся в БД, схема которой описывается в пункте 2.6.

Для работы правил необходима информация о двух СГ, составляющих объекты применения правила, о том, какая из двух СГ – главная, а какая – подчиненная. Также необходима информация о самом приоритете правила и о флагах согласования двух СГ по роду, числу и падежу.

Основной задачей в процессе изучения предметной области является выявления зависимостей и закономерностей формирования синтаксических конструкций.

В основе синтаксического анализа выбранного подхода лежит построение синтаксических групп на одном морфологическом варианте разбора части предложения. Частью называется простое предложение, которое содержит только одну грамматическую основу. Перед непосредственно анализом одного фрагмента производится фрагментация сложного предложения на простые.

Существует два типа синтаксического анализа:

* Нисходящий синтаксический анализ. При нисходящем анализе делается предположение о строении предложения и проверяется путем сравнения с исходным. Таким образом, движение идет от корня к листьям.
* Восходящий синтаксический анализ. При восходящем синтаксическом анализе с помощью набора правил единицы предложения объединяются в группы, пока в предложении не останется одна группа, которая будет соответствовать корню – грамматической основе.

Для того, чтобы унифицировать применение различных правил необходимо заранее разработать структуру, которая будет учитывать все возможные варианты и разработать алгоритм применения правил, соответствующих данной структуре. Слово рассматривается исключительно как СГ, и не важно, сколько слов она подчинила и подчинила ли. Каждое словосочетание (СГ) выполняет определенную синтаксическую функцию, которая определяется главным словом в данном словосочетании. Словосочетания могут объединятся в более крупные, сохраняя роль главного или приобретая новую.

Синтаксическая группа (СГ) может обладать синтаксической ролью, также в ней определено главное слово, которое определяет роль словосочетания. В таком случае и слова, и словосочетания будут является синтаксическими группами.

На вход алгоритма подаются две синтаксические группы, по которым делается запрос в таблицу БД «Правило». Алгоритм определяет уникальное правило сочетания именно этих двух СГ. Далее по приоритету выбранного правила определяется, необходимо ли применять данное правило (подробное описание алгоритма анализа с применением правил приводится в пункте 2.7.2). Затем правило применяется, и из двух СГ остается доступной для анализа только одна из них: главная.

Всего у входных СГ может быть два типа: главная и подчиненная.

В русском языке есть правила сочетания слов, как правило, не закрепленные документально. В русском языке нет правил порядка следования членов предложения в связи с чем моментально и многократно возрастает сложность нахождения отношений между словами. Однако с этим ничего не поделать, поэтому необходимо просто обработать в два-три раза больше вариантов сочетаний слов, чем для других языков.

Отношения СГ, по сути, отношения зависимости с некоторыми оговорками, поскольку возможно нахождение однородных членов и даже однородных простых предложений в составе сложного. Однако правила устанавливают именно отношения зависимости, выстраивая иерархию слов предложения.

Само правило хранит свой приоритет для определения порядка применения правил. Также правило содержит информации о необходимом согласовании входных групп, которые к нему относятся. Это позволяет обеспечивать требования согласования в различных правилах по роду, числу и падежу.

Все правила однозначно определяют итоговый член предложения, который реализует подчиненная СГ данного словосочетания.

Как говорилось ранее у каждого правила есть приоритет. Этот приоритет соответствует порядку построения групп: от меньших к большим. Например, сначала надо построить группы нареч-ПРИЛ, а потом прил-СУЩ, чтобы построить структуру на отрезке "очень красивый человек": прил-СУЩ (нареч-ПРИЛ (очень красивый), человек) Каждое правило применяется к каждому слову входного отрезка слева направо.

Ниже представлены примеры правил и их описания в рамках входных синтаксических групп.

На рисунке 4 приведена структура правила «наречие + глагол».

* 1. Проектирование модели базы данных

Для создания оптимальной базы данных необходимо пройти следующие три этапа:

* Концептуальное проектирование БД, т.е. построение инфологической модели БД;
* Логическое проектирование БД, т.е. построение даталогической модели БД;
* Физическое проектирование БД, т.е. построение самой БД средствами выбранной СУБД.

Основной задачей на этапе концептуального проектирования БД является выделение объектов данных, их свойств и связей между ними. Построение оптимальной инфологической модели позволит минимизировать возможность создания лишних, или отсутствия необходимых объектов и связей, замедляющих процесс обработки данных при работе разрабатываемого ПО. Основными понятиями на данном этапе являются: объекты, сущности, параметры, атрибуты.

Объект – элемент действительности, обладающий набором параметров, т.е. свойств. Отображение какого-либо объекта называется сущностью. Параметры объекта хранятся при сущности в виде её атрибутов.

В базе данных сущность становится таблицей. В таблице, как всегда, есть строки и столбцы. Одна строка называется записью. Запись – конкретный экземпляр данной сущности, обладающий тем же набором атрибутов, что и другие записи в таблице. Поля записи содержат значения атрибутов. Таким образом столбец – множество значений одного атрибута у всех записей из таблицы. Среди атрибутов выделяется один или несколько, которые называются первичным ключом, если для них выполняется условие уникальности, позволяющее однозначно определить каждую конкретную запись.

Связи отражают отношения между сущностями, и реализуются с помощью внешних ключей, т.е. полей первичных ключей, перешедших из других таблиц.

На этапе концептуального проектирования строится модель «Сущность-Связь», или ER-модель, или Инфологическая модель. Используются следующие графические средства: сущность, атрибут, связь.

По кардинальности связи классифицируют на три вида: «один-к-одному», «один-ко-многим», «многие-ко-многим». Связь «один к одному» строится тогда, когда для двух сущностей выполняется взаимное условие: один экземпляр одной сущности может быть связан с не более чем одним экземпляром другой сущности. Связь «один-ко-многим» строится, когда для двух сущностей выполняется условие: один экземпляр первой сущности может быть связан с несколькими экземплярами второй сущности, а один экземпляр второй сущности может быть связан с не более чем одним экземпляром первой сущности. Связь «многие-ко-многим» строится, когда для двух сущностей выполняется взаимное условие: один экземпляр одной сущности может быть связан с несколькими экземплярами другой сущности.

Каждой связи присваивается имя, которое отражает семантику взаимодействия экземпляров сущностей.

Прежде чем строить Инфологическую модель, необходимо построить таблицы сущностей будущей БД с указанием типов атрибутов и указанием ключевых атрибутов.

Ключевая сущность БД: «Правило». Она необходима для сохранения информации обо всех применяемых правилах, касающихся выделения синтаксических конструкций простого предложения. Именно эти правила сочетаемости словоформ применяются на этапе синтаксического анализа.

Сущность «Правило» представлена в таблице 1.

**Таблица 1. Сущность «Правило»**

| **Логическое имя атрибута** | **Тип данных** | **Ключ** |
| --- | --- | --- |
| ID правила | Целочисленный | Первичный (PK) |
| Название | Текстовый |  |
| Приоритет | Числовой |  |
| Флаг согласования рода | Логический |  |
| Флаг согласования числа | Логический |  |
| Флаг согласования падежа | Логический |  |

Ниже в таблицах 1-2 представлены две сущности, которые отвечают за сохранение информации о примененных правилах и тех словах, к которым эти правила были применены. В связи с выбранным алгоритмом синтаксического анализа, необходимо сохранить информацию о главной и подчиненной СГ.

Сущность «Главная синтаксическая группа» представлена в таблице 2.

**Таблица 2. Сущность «Главная синтаксическая группа»**

| **Логическое имя атрибута** | **Тип данных** | **Ключ** |
| --- | --- | --- |
| ID главной группы | Целочисленный | Первичный (PK) |
| Содержимое группы | Текстовый |  |
| Приоритет | Числовой |  |
| Номер предложения | Числовой |  |
| Номер части | Числовой |  |
| Номер варианта | Числовой |  |
| Номер в части | Числовой |  |

Сущность «Подчиненная синтаксическая группа» представлена в таблице 3.

**Таблица 3. Сущность «Подчиненная синтаксическая группа»**

| **Логическое имя атрибута** | **Тип данных** | **Ключ** |
| --- | --- | --- |
| ID подчиненной группы | Целочисленный | Первичный (PK) |
| Содержимое группы | Текстовый |  |
| Приоритет | Числовой |  |
| Номер предложения | Числовой |  |
| Номер части | Числовой |  |
| Номер варианта | Числовой |  |
| Номер в части | Числовой |  |

Сущность «Тип входной синтаксической группы» представлена в таблице 4. Она описывает два возможных типа: «подчиненный» и «главный», в соответствии с которыми идёт управление двумя СГ, которые связываются по какому-либо правилу. Таким образом типы 1-й и 2-й СГ определяется правилом.

**Таблица 4. Сущность «Тип входной синтаксической группы»**

| **Логическое имя атрибута** | **Тип данных** | **Ключ** |
| --- | --- | --- |
| ID типа входной синтаксической группы | Целочисленный | Первичный (PK) |
| Название типа входной синтаксической группы | Текстовый |  |

Сущность «Вид синтаксической группы» представлена в таблице 5. Данная сущность унифицирует синтаксические группы по видовому признаку, а именно их главные слова, благодаря чему делается возможным применение правила к двум СГ.

**Таблица 5. Сущность «Вид синтаксической группы»**

| **Логическое имя атрибута** | **Тип данных** | **Ключ** |
| --- | --- | --- |
| ID вида синтаксической группы | Целочисленный | Первичный (PK) |
| Название вида синтаксической группы | Текстовый |  |

Сущность «Часть речи» представлена в таблице 6. Каждое слово – представитель определенной части речи русского языка, поэтому эта характеристика – обязательная связь между русским языком и принципом СГ.

**Таблица 6. Сущность «Часть речи»**

| **Логическое имя атрибута** | **Тип данных** | **Ключ** |
| --- | --- | --- |
| ID части речи | Целочисленный | Первичный (PK) |
| Название части речи | Текстовый |  |

Сущность «Граммема» представлена в таблице 7.

**Таблица 7. Сущность «Граммема»**

| **Логическое имя атрибута** | **Тип данных** | **Ключ** |
| --- | --- | --- |
| ID граммемы | Целочисленный | Первичный (PK) |
| Содержание граммемы | Текстовый |  |

Сущность «Класс синтаксической группы» представлена в таблице 8.

**Таблица 8. Сущность «**Класс синтаксической группы**»**

| **Логическое имя атрибута** | **Тип данных** | **Ключ** |
| --- | --- | --- |
| ID класса | Целочисленный | Первичный (PK) |
| Название класса | Текстовый |  |

Сущность «Член предложения» представлена в таблице 9. Данная таблица дает информацию о том, какой член предложения реализуется той или иной СГ.

**Таблица 9. Сущность «Член предложения»**

| **Логическое имя атрибута** | **Тип данных** | **Ключ** |
| --- | --- | --- |
| ID члена предложения | Целочисленный | Первичный (PK) |
| Название члена предложения | Текстовый |  |

Ниже представлены три сущности, соответствующие трём видам граммем русского языка: род, число и падеж, в соответствии с которыми определяется лексически правильное взаимодействие слов.

Сущность «Род» представлена в таблице 10.

**Таблица 10. Сущность «Род»**

| **Логическое имя атрибута** | **Тип данных** | **Ключ** |
| --- | --- | --- |
| ID рода | Целочисленный | Первичный (PK) |
| Название рода | Текстовый |  |

Сущность «Число» представлена в таблице 11.

**Таблица 11. Сущность «Число»**

| **Логическое имя атрибута** | **Тип данных** | **Ключ** |
| --- | --- | --- |
| ID числа | Целочисленный | Первичный (PK) |
| Название числа | Текстовый |  |

Сущность «Падеж» представлена в таблице 12.

**Таблица 12. Сущность «Падеж»**

| **Логическое имя атрибута** | **Тип данных** | **Ключ** |
| --- | --- | --- |
| ID падежа | Целочисленный | Первичный (PK) |
| Название падежа | Текстовый |  |

Ключевые атрибуты, показанные в таблицах 1-12, на Инфологической модели БД выделяются подчёркиванием, но прежде, чем строить эту модель, необходимо определить связи между сущностями БД. А на основе ключевых атрибутов в дальнейшем будут строится первичные ключи в таблицах Даталогической модели.

Преобразование ER-модели в реляционную модель БД происходит с учётом типа связи:

* Связь «1:1». Для реализации данного типа связи требуется от одной до двух таблиц в зависимости от того, обязательно ли поле.
* Связь «1:М». Связь может быть реализована двумя или тремя таблицами.
* Связь «М:1». Также необходимо две или три таблицы.
* Связь «М:N». Такая связь требует обязательного построения третей таблицы – одну для каждой сущности и одну для связи.

Первичный ключ должен удовлетворять ряду требований:

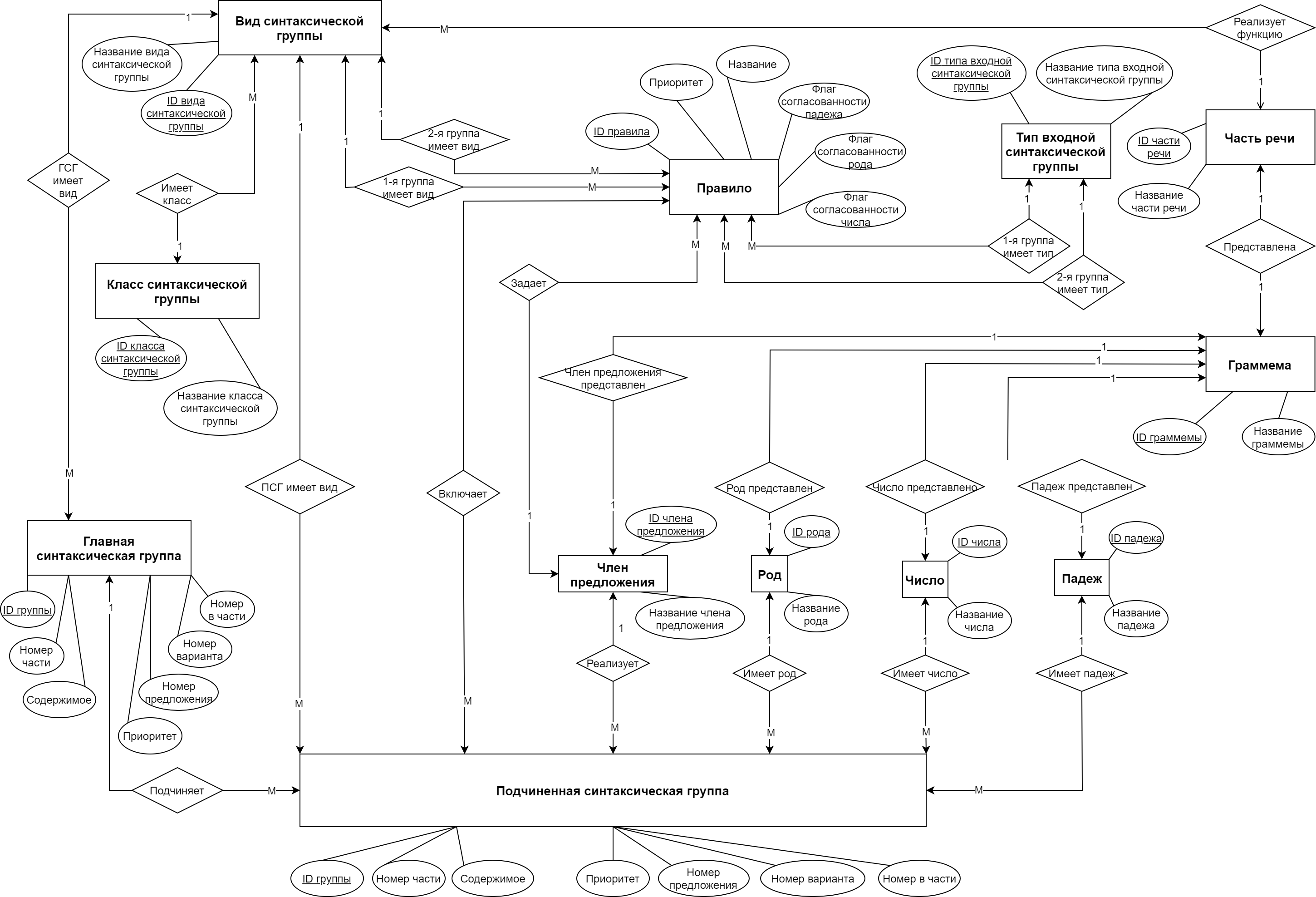
* Уникальность. Два экземпляра не должны иметь одинаковых значений ключа.
* Минимум атрибутов. Составной первичный ключ (группа атрибутов) не должен содержать ни одного атрибута, удаление которого не приводило бы к утрате уникальности.

Внешний ключ (Foreign Key) появляется в процессе установления связей между сущностями. Это описательный атрибут, который назначается в дополнение к идентифицирующему атрибуту и служит не для идентификации этого уникального экземпляра, а для выделения набора интересующих тех объектов по каким-либо внешним признакам, привязанным к этим объектам из других таблиц.

С учётом описания сущностей формируются следующие связи:

* «Род представлен»: связь между сущностями «Граммема» и «Род» типа «один-к-одному»
* «Число представлено»: связь между сущностями «Граммема» и «Число» типа «один-к-одному»
* «Падеж представлен»: связь между сущностями «Граммема» и «Падеж» типа «один-к-одному»
* «Часть речи представлена»: связь между сущностями «Граммема» и «Часть речи» типа «один-к-одному»
* «Член предложения представлен»: связь между сущностями «Граммема» и «Член предложения» типа «один-к-одному»
* «Реализует функцию»: связь между сущностями «Часть речи» и «Вид синтаксической группы» типа «один-ко-многим»
* «1-я группа имеет тип»: связь между сущностями «Тип входной синтаксической группы» и «Правило» типа «один-ко-многим»
* «2-я группа имеет тип»: связь между сущностями «Тип входной синтаксической группы» и «Правило» типа «один-ко-многим»
* «1-я группа имеет вид»: связь между сущностями «Вид синтаксической группы» и «Правило» типа «один-ко-многим»
* «2-я группа имеет вид»: связь между сущностями «Вид синтаксической группы» и «Правило» типа «один-ко-многим»
* «Имеет класс»: связь между сущностями «Класс синтаксической группы» и «Вид синтаксической группы» типа «один-ко-многим»
* «ГСГ имеет вид»: связь между сущностями «Вид синтаксической группы» и «Главная синтаксическая группа» типа «один-ко-многим» (ГСГ – Главная Синтаксическая Группа)
* «ПСГ имеет вид»: связь между сущностями «Вид синтаксической группы» и «Подчиненная синтаксическая группа» типа «один-ко-многим» (ПСГ – Подчиненная Синтаксическая Группа)
* «Определена»: связь между сущностями «Правило» и «Подчиненная синтаксическая группа» типа «один-ко-многим»
* «Подчиняется»: связь между сущностями «Вид синтаксической группы» и «Подчиненная синтаксическая группа» типа «один-ко-многим»
* «Реализует»: связь между сущностями «Член предложения» и «Подчиненная синтаксическая группа» типа «один-ко-многим»
* «Задает»: связь между сущностями «Член предложения» и «Правило» типа «один-ко-многим»
* «Имеет род»: связь между сущностями «Род» и «Подчиненная синтаксическая группа» типа «один-ко-многим»
* «Имеет число»: связь между сущностями «Число» и «Подчиненная синтаксическая группа» типа «один-ко-многим»
* «Имеет падеж»: связь между сущностями «Падеж» и «Подчиненная синтаксическая группа» типа «один-ко-многим».

На рисунке 1 представлена Инфологическая модель БД.

Рисунок 1. Схема Инфологической модели.

Под даталогической понимается модель, отражающая логические взаимосвязи между элементами данных безотносительно их содержания и физической организации. Даталогическое проектирование включает построение БД в виде таблиц. Особенности выбранной СУБД также влияют на проеткирование даталогической модели, например, при описании названий типов данных, их размера. Результатом данной стадии проектирования БД является построение датологической модели в виде набора таблиц, отражающих основные свойства и взаимодействия сущностей предметной области.

Таблица содержит название и набор полей, которые характеризуются типом данных. Поле может содержать информативные данных стандартных типов данных (целый, с плавающей точкой, строковый и т.д.) или быть ссылкой на другую таблицу – внешним ключом.

При добавлении информации в БД в таблицы вставляются записи, для которых каждое поле принимает значение заданного типа.

Проектированию базы данных уделяется особое внимание по причине того, что внесение изменений в структуру при наличии данных влечет дополнительные сложности, объясняющиеся необходимостью специальными методами перезаписывать и изменять все имеющиеся данные.

Данная предметная область облегчает построение Базы данных, поскольку ее данные используют лишь алгоритмы синтаксического анализатора. Она не будет содержать конфиденциальных или сколь угодно уникальных данных, поэтому БД проекта не нуждается в дополнительных средствах защиты данных, не предусмотренных базовыми настройками выбранной СУБД.

Модель «Rule» представлена в таблице 13.

**Таблица 13. Модель «Rule»**

| **Логическое имя атрибута** | **Поле** | **Тип данных** | **Default** | **Ключ** |
| --- | --- | --- | --- | --- |
| ID правила | ID | INT(11) | Auto Incremental | Primary Key |
| Название правила | Name | VARCHAR(45) | NULL |  |
| Приоритет | Priority | INT(11) | NULL |  |
| Флаг согласования рода | Gender consistency | TINYINT(4) | NULL |  |
| Флаг согласования числа | Number consistency | TINYINT(4) | NULL |  |
| Флаг согласования падежа | Case consistency | TINYINT(4) | NULL |  |
| Вид первой СГ | 1G Syntax kind ID | INT(11) | NULL | Foreign Key |
| Вид второй СГ | 1G Syntax kind ID | INT(11) | NULL | Foreign Key |
| Тип первой СГ | 1G Syntax kind ID | INT(11) | NULL | Foreign Key |
| Тип второй СГ | 1G Syntax kind ID | INT(11) | NULL | Foreign Key |
| Член предложения подчиненной группы | 1G SG attribute ID | INT(11) | NULL | Foreign Key |

Модель «Syntax kind» представлена в таблице 14.

**Таблица 14. Модель «Syntax kind»**

| **Логическое имя атрибута** | **Поле** | **Тип данных** | **Default** | **Ключ** |
| --- | --- | --- | --- | --- |
| ID вида СГ | ID | INT(11) | Auto Incremental | Primary Key |
| Название вида СГ | Name | VARCHAR(45) | NULL |  |
| Части речи | Part of speech ID | INT(11) | NULL | Foreign Key |
| Класс СГ | Syntax class ID | INT(11) | NULL | Foreign Key |

Модель «Input syntax type» представлена в таблице 15.

**Таблица 15. Модель «Input syntax type»**

| **Логическое имя атрибута** | **Поле** | **Тип данных** | **Default** | **Ключ** |
| --- | --- | --- | --- | --- |
| ID правила | ID | INT(11) | Auto Incremental | Primary Key |
| Название правила | Name | VARCHAR(45) | NULL |  |

Модель «Attribute» представлена в таблице 16.

**Таблица 16. Модель «Attribute»**

| **Логическое имя атрибута** | **Поле** | **Тип данных** | **Default** | **Ключ** |
| --- | --- | --- | --- | --- |
| ID члена предложения | ID | INT(11) | Auto Incremental | Primary Key |
| Название члена предложения | Name | VARCHAR(45) | NULL |  |

Модель «Syntax class» представлена в таблице 17.

**Таблица 17. Модель «Syntax class»**

| **Логическое имя атрибута** | **Поле** | **Тип данных** | **Default** | **Ключ** |
| --- | --- | --- | --- | --- |
| ID класса СГ | ID | INT(11) | Auto Incremental | Primary Key |
| Название класса СГ | Name | VARCHAR(45) | NULL |  |

Модель «Grammeme» представлена в таблице 18.

**Таблица 18. Модель «Grammeme»**

| **Логическое имя атрибута** | **Поле** | **Тип данных** | **Default** | **Ключ** |
| --- | --- | --- | --- | --- |
| ID граммемы | ID | INT(11) | Auto Incremental | Primary Key |
| Название граммемы | Name | VARCHAR(45) | NULL |  |

Модель «Gender» представлена в таблице 19.

**Таблица 19. Модель «Gender»**

| **Логическое имя атрибута** | **Поле** | **Тип данных** | **Default** | **Ключ** |
| --- | --- | --- | --- | --- |
| ID рода | ID | INT(11) | Auto Incremental | Primary Key |
| Название рода | Name | VARCHAR(45) | NULL |  |
| Граммема | Grammeme ID | INT(11) | NULL | Foreign Key |

Модель «Case» представлена в таблице 20.

**Таблица 20. Модель «Case»**

| **Логическое имя атрибута** | **Поле** | **Тип данных** | **Default** | **Ключ** |
| --- | --- | --- | --- | --- |
| ID падежа | ID | INT(11) | Auto Incremental | Primary Key |
| Название падежа | Name | VARCHAR(45) | NULL |  |
| Граммема | Grammeme ID | INT(11) | NULL | Foreign Key |

Модель «Number» представлена в таблице 21.

**Таблица 21. Модель «Number»**

| **Логическое имя атрибута** | **Поле** | **Тип данных** | **Default** | **Ключ** |
| --- | --- | --- | --- | --- |
| ID числа | ID | INT(11) | Auto Incremental | Primary Key |
| Название числа | Name | VARCHAR(45) | NULL |  |
| Граммема | Grammeme ID | INT(11) | NULL | Foreign Key |

Модель «Part of speech» представлена в таблице 22.

**Таблица 22. Модель «Part of speech»**

| **Логическое имя атрибута** | **Поле** | **Тип данных** | **Default** | **Ключ** |
| --- | --- | --- | --- | --- |
| ID часть речи | ID | INT(11) | Auto Incremental | Primary Key |
| Название части речи | Name | VARCHAR(45) | NULL |  |
| Граммема | Grammeme ID | INT(11) | NULL | Foreign Key |

Модель «Main syntax group» представлена в таблице 23.

**Таблица 23. Модель «Main syntax group»**

| **Логическое имя атрибута** | **Поле** | **Тип данных** | **Default** | **Ключ** |
| --- | --- | --- | --- | --- |
| ID группы | ID | INT(11) | Auto Incremental | Primary Key |
| Содержимое | Content | VARCHAR(45) | NULL |  |
| Приоритет | Priority | INT(11) | NULL |  |
| Номер предложения | Sentence number | INT(11) | NULL |  |
| Номер части | Part number | INT(11) | NULL |  |
| Номер варианта | Variant number | INT(11) | NULL |  |
| Номер в части | Number in part | INT(11) | NULL |  |
| Вид СГ | Syntax kind ID | INT(11) | NULL | Foreign Key |

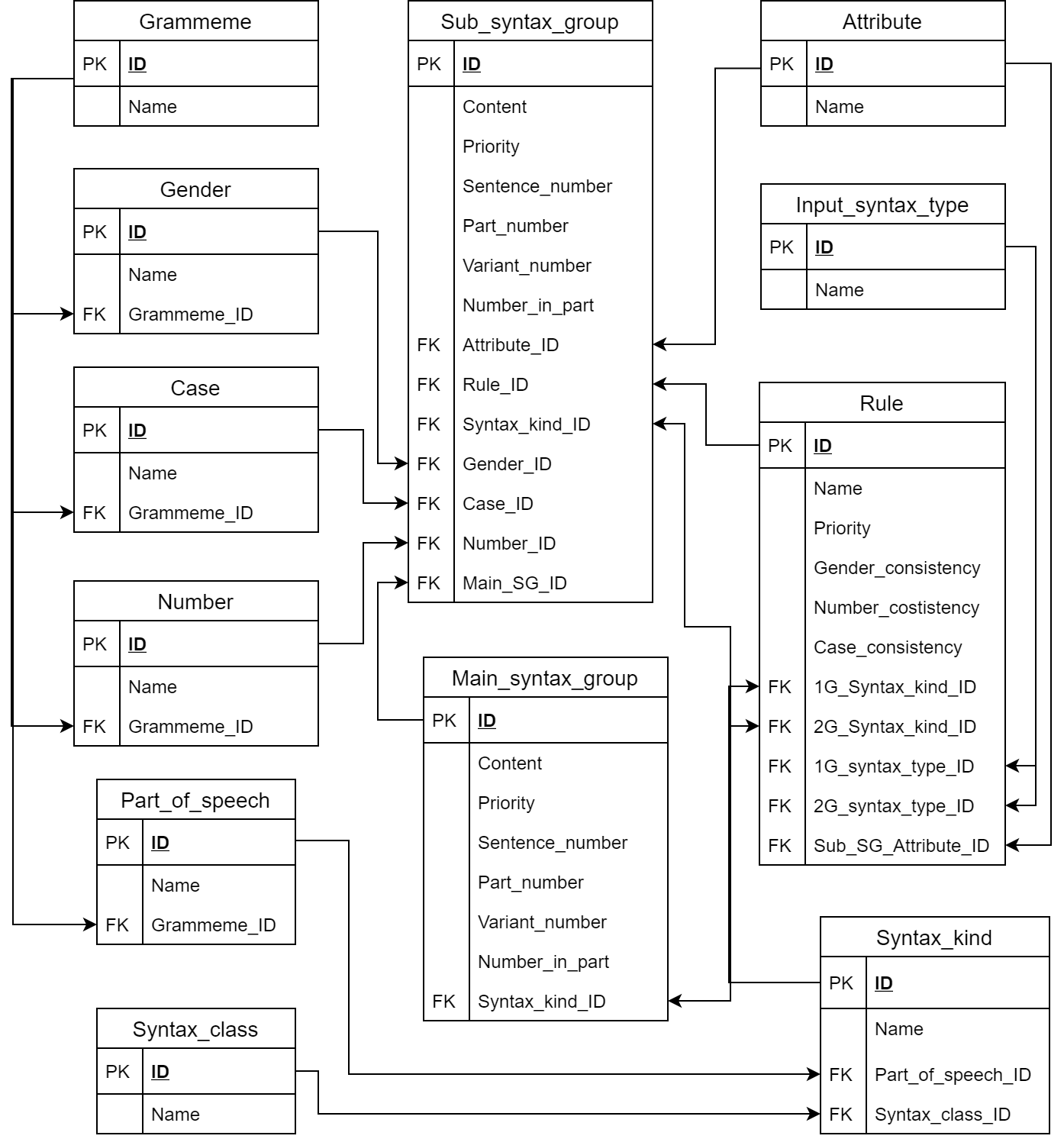
Модель «Sub syntax group» представлена в таблице 24.

**Таблица 24. Модель «Sub syntax group»**

| **Логическое имя атрибута** | **Поле** | **Тип данных** | **Default** | **Ключ** |
| --- | --- | --- | --- | --- |
| ID группы | ID | INT(11) | Auto Incremental | Primary Key |
| Содержимое | Content | VARCHAR(45) | NULL |  |
| Приоритет | Priority | INT(11) | NULL |  |
| Номер предложения | Sentence number | INT(11) | NULL |  |
| Номер части | Part number | INT(11) | NULL |  |
| Номер варианта | Variant number | INT(11) | NULL |  |
| Номер в части | Number in part | INT(11) | NULL |  |
| Член предложения | Attribute ID | INT(11) | NULL | Foreign Key |
| Правило | Rule ID | INT(11) | NULL | Foreign Key |
| Вид СГ | Syntax kind ID | INT(11) | NULL | Foreign Key |
| Род | Gender ID | INT(11) | NULL | Foreign Key |
| Число | Number ID | INT(11) | NULL | Foreign Key |
| Падеж | Case ID | INT(11) | NULL | Foreign Key |
| Главная СГ | Main SG ID | INT(11) | NULL | Foreign Key |

Связи даталогической модели формируются на основе предыдущих этапов проектирования. На рисунке 2 представлена итоговая схема Даталогической модели.

По результатам данного этапа проектирования строится сама БД средствами СУБД MySQL.

Рисунок 2. Схема даталогической модели

* 1. Разработка программного продукта
     1. Разработка интерфейса пользователя

Партия

* + 1. Разработка алгоритмов синтаксического анализа текста

После получения морфологической информации (с помощью библиотек ru\_sent\_tokenize, NLTK и pymorphy2) о составе предложения начинается этап синтаксического анализа, который проходит в два этапа:

1. Производится разделение сложных предложений на простые:

* Формирование списка союзов, которые могут разделять простые предложения в составе сложного (сложно-подчиненные и сложно-сочиненные). Прежде чем приступать к следующей стадии, происходит фильтрование ложных вариантов союзов: например, были найдены варианты союза (между элементами которого нет других союзов) «если …, то»: «если …, то» и «если …, то …, то» – в этом случае будет выбран второй вариант, поскольку отдельного союза «то» не существует, а сочетание «, то», где «то» - не часть союза, не является корректной для русского языка;
* С помощью оптимального списка союзов-разделителей (в котором, например, составные союзы не разделяются другими союзами) происходит выделение всех возможных пар подряд идущих простых предложений в составе сложного;
* Для каждой пары производится проверка наличия вариантов грамматических основ двух возможных частей предложения, которые разделяет какой-то союз-разделитель. Две части в составе сложного предложения формируются только в том случае, если в обеих этих частях есть свои варианты грамматической основы. Если хотя бы в одной из этих двух предполагаемых частей грамматическая основа не может быть сформирована, то эти две части не разделяются.

2) Анализ каждой части предложения с учётом всех возможных вариантов грамматической основы данной части:

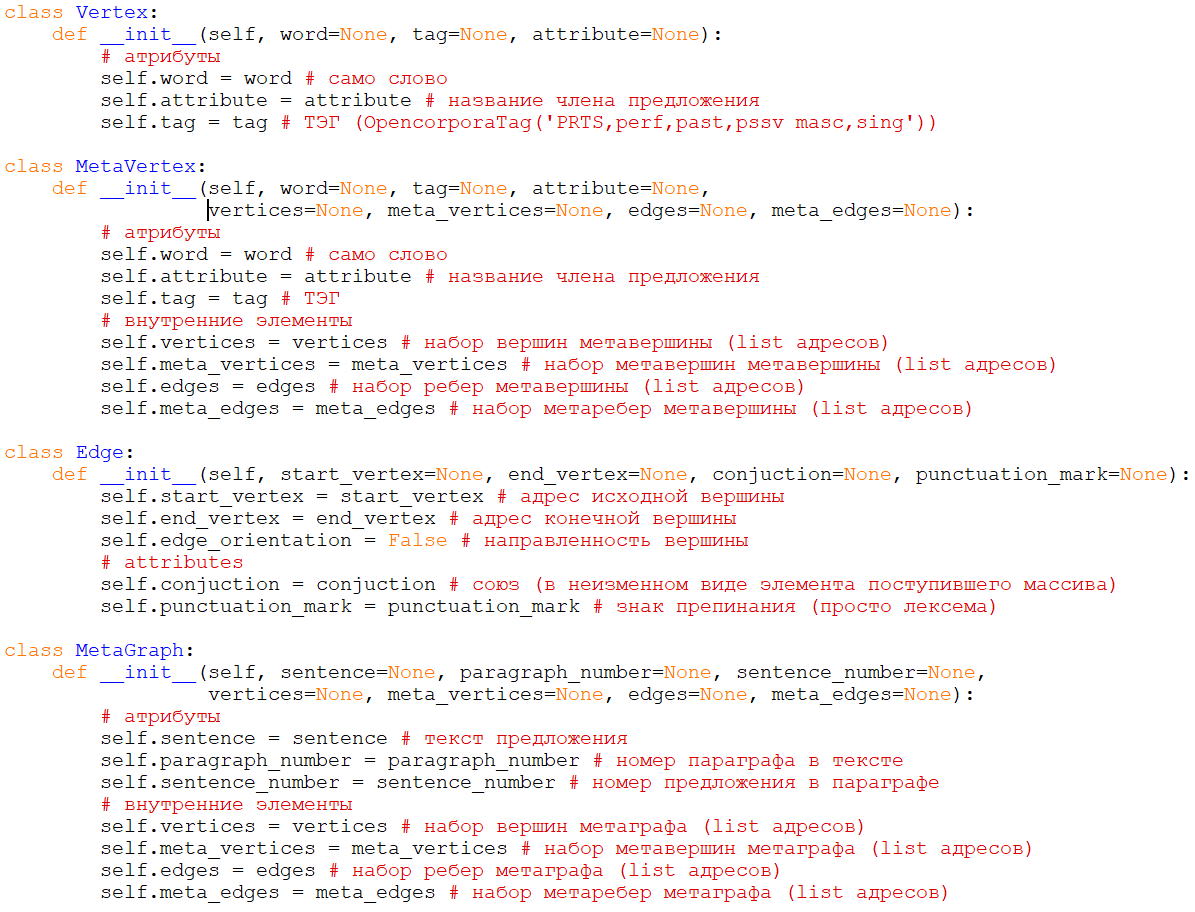
* Так как каждая лексема части может иметь несколько вариантов морфологического разбора, с помощью рекурсивного алгоритма формируется необходимое количество (произведение морфологических вариантов всех лексем части) уникальных цепочек из вариантов лексем;
* Каждый такой вариант цепочки рассматривается вслед за предыдущим на следующей стадии;
* Если этот вариант сохранил хотя бы один из вариантов грамматической основы, которую изначально могла иметь данная часть после шага 1), производится синтаксический анализ с использованием правил (подробно схема применения правил описана в пункте 2.5);
* Вариант считается успешно разобранным, если после синтаксического анализа не осталось неразобранных слов, после этого он сохраняется в списке вариантов разбора части данного предложения;
  + 1. Разработка структуры классов модуля построения метаграфа

В своей работе я использую метаграфовую модель в соответствии с источником [3]:

где MG – метаграф, V – множество вершин метаграфа; MV – множество метавершин метаграфа; E – множество ребер метаграфа; ME – множество метаребер метаграфа.

Единственным исключением в применении данной модели на этапе синтаксического анализатора является то, что не используются метаребра. Функционала Вершин, Рёбер и Метавершин достаточно для реализации записей о словах и об их взаимодействии между собой.

На рисунке 4 представлен код Классов Вершина, Метавершина, Ребро и Метаграф соответственно.

Рис. 4. Классы метаграфа и его элементов

У вершины-слова атрибутами являются: само слово, член предложения, который реализует слово, а также тэг слова. Тэг - набор морфологических характеристик, которые имеет данное слово по правилам русского языка, и взятых из словаря OpenCorpora. Данная информация о конкретном слове, а также значение поля attribute являются необходимой и достаточной информации о лексической роли слова в предложении. Полный список тэгов вы можете посмотреть в источнике [6], являющимся страницей официального сайта OpenCorpora.

У метавершины-части предложения атрибутами являются те же поля. Значением атрибута word в нем будет являться грамматическая основа предложения в формате «подлежащее\_сказуемое». Также метавершина имеет набор вершин, ребер, метавершин, которые могут входить в её состав.

У ребра атрибуты: исходная и конечная вершины, а также ориентация вершины, союз и знак пунктуации. Последние два атрибута необходимы, чтобы сохранить информацию о связующих лексемах предложения: союзах и знаках препинания: запятая, двоеточие, тире. Знаки препинания несут определенную семантическую нагрузку, о которой мы не можем делать выводы на данном синтаксическом этапе разбора предложения.

Метаграф - это хранилище информации о предложении, поэтому одним из его атрибутов является само предложение. Также атрибутами являются номер параграфа, в котором находится данное предложение и номер предложения в данном абзаце. Объект метаграфа хранит список вершин, метавершин, ребер и метарёбер данного метаграфа.

* + 1. Разработка алгоритма преобразования результатов синтаксического анализа в метаграф

Так как предложения это, по сути, набор слов, то помимо самого слова, метавершина должна обладать информацией о том, какой номер имеет данное слово в данной части предложения. Таким образом слова внутри метавершины следует хранить не как список, а как словарь. Элементом словаря будет являться само слово, а ключом к данному элементу – его номер в предложении.

Таким же образом хранится информация о рёбрах, находящихся в составе метавершины. Ключом к элементу ребра является кортеж из двух значений: номера главного и подчинённого слов в части предложения.

Информация о рёбрах, которые связывают метавершины метаграфа, тоже хранится в словаре. Ключом к элементу данного словаря является кортеж из 4 элементов: номера главной и подчиненной частей предложения, номера вариантов разбора главной и подчиненной частей предложения. Таким образом, какую бы главную и подчиненную вершины, мы бы не связали, какой бы вариант разбора части предложения мы не выбрали, мы получим уникальный ключ для каждого ребра. Данный ключ позволит получить доступ к уникальным метавершинам и вершинам, информация о которых не будет дублироваться, благодаря структурам данных языка Python 3.

У каждой части предложения возможно несколько вариантов разбора, которые никак не связаны между собой. Поэтому каждый такой вариант разбора будет являться элементом соответствующего списка метавершин. Каждый элемент данного списка является в принципе равноправным по сравнению с другими элементами после осуществления синтаксического разбора. Тот или иной вариант разбора части получает больший номер в связи с результатами морфологического и синтаксического разборов, которые помимо всего прочего показывают вероятность актуальности того или иного варианта: чем больше вероятность, тем больше номер варианта в списке. Таким образом, метаграф хранит словарь метавершин, элементами которых являются списки вариантов разбора каждой части предложения.

Таким образом мы храним информацию без излишеств и в максимально удобной и формализованной форме без хранения в памяти лишних индексов и обеспечением максимально быстрого доступа к этим данным. В будущем возможна разработка модуля, который позволит записывать информацию о разборе метаграфа в специальный файл, возможно, уникального формата.

* 1. Результаты опытной эксплуатации

Общие требования к анализируемым системам:

ЗАКЛЮЧЕНИЕ

При разработке информационной системы планирования и учета личных тренировок были получены следующие результаты:

* Была изучена предметная область, связанная ххх ххххх хххххх ххххххх хххххххх хххххххх ххх ххххх хххххх ххххххх ххххххххх ххххххххх ххххххххх ххххххх ххххххх ххххх хххх ххх ххххх хх ххххххххххх ххх ххххххх хххххх хххх хххххххххх ххххх ххххх хх хх хххххххххххх ххх ххххх хххххххххх хх ххххх хххххх ххх хххххххх ххххх ххх ххххх ххххххх.
* Были сформулированы требования ххх ххххх хххххх ххххххх хххххххх хххххххх ххх ххххх хххххх ххххххх ххххххххх ххххххххх ххххххххх ххххххх ххххххх.
* Спроектирована ххх ххххх хххххх ххххххх хххххххх хххххххх ххх ххххх хххххх ххххххх ххххххххх ххххххххх ххххххххх ххххххх ххххххх ххххх.
* Разработана ххх ххххх хххххх ххххххх хххххххх хххххххх ххх ххххх хххххх ххххххх ххххххххх ххххххххх ххххххххх ххххххх ххххххх ххххх хххх ххх.

Полученное информационно-программное изделие имеет возможности расширения за счет подключения новых платформ к серверному приложению. Это позволит системе привлечь новых пользователей и оставаться конкурентно способной среди аналогов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Волкова, И. А. Введение в компьютерную лингвистику. Практические аспекты создания лингвистических процессоров: Учебное пособие для студентов ВМиК МГУ / И. А. Волкова – Москва: Издательский отдел факультета ВМиК МГУ, 2006. – 43 с.
2. Прикладная и компьютерная лингвистика / ред. Николаев И. С., Митренина О. В., Ландо Т. М. – 2-е изд., – Москва: Издательская группа URSS, 2017. – 320 с.
3. Гапанюк, Ю. Е. Конспект лекций по спецкурсу «Гибридные интеллектуальные информационные системы на основе метаграфового подхода» / Ю. Е. Гапанюк – Москва: Издательство «Спутник +», 2018. – 56 с.
4. Тузов, В. А. Компьютерная семантика русского языка / В. А. Тузов – СПб.: Изд-во СПбГУ, 2003. – 391 с.
5. Мартынов, В. В. Основы семантического кодирования. Опыт Представления и преобразования знаний / В. В. Мартынов. – Минск: ЕГУ, 2001 г. – 140 с.
6. [Электронный ресурс] – 2009-2020 г. – Режим доступа: http://opencorpora.org/dict.php?act=gram, свободный.
7. [Электронный ресурс] / М. Коробов – 2015-2020 г. – Режим доступа: https://pymorphy2.readthedocs.io/en/latest/index.html, свободный
8. Постников, В. М. Основы эксплуатации АСОИиУ Том 1. Техническое обслуживание АСОИиУ. 2-е издание, переработанное и дополненное, учеб. пособие / В. М. Постников. – М.: Изд-во МГТУ им. Н. Э. Баумана, 2015. – 214 [2] c.: ил.

ПРИЛОЖЕНИЕ А   
ГРАФИЧЕСКИЕ МАТЕРИАЛЫ

1. Цели и задачи

2. Предметная область

3. Сравнение с аналогами

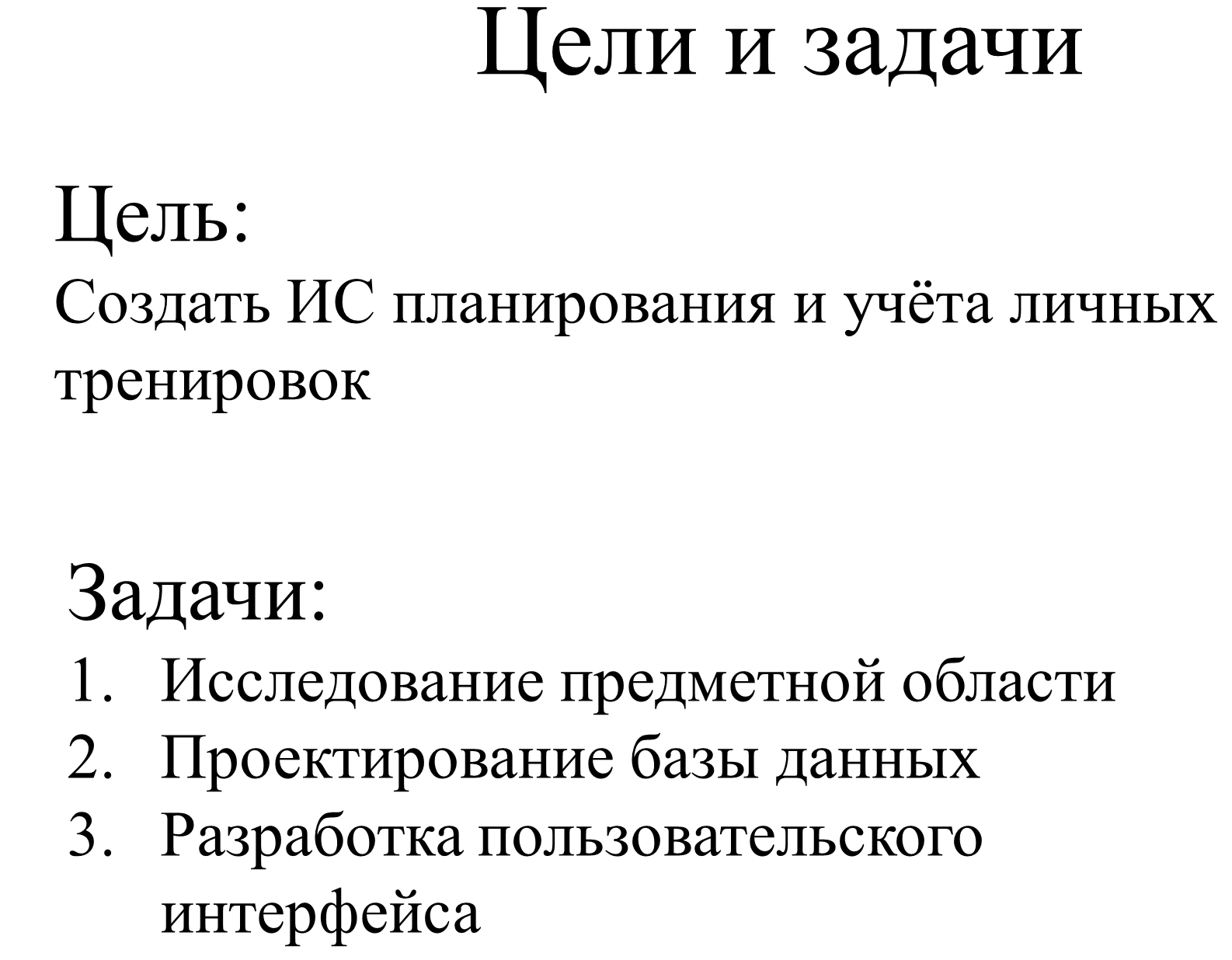
4. Архитектура системы

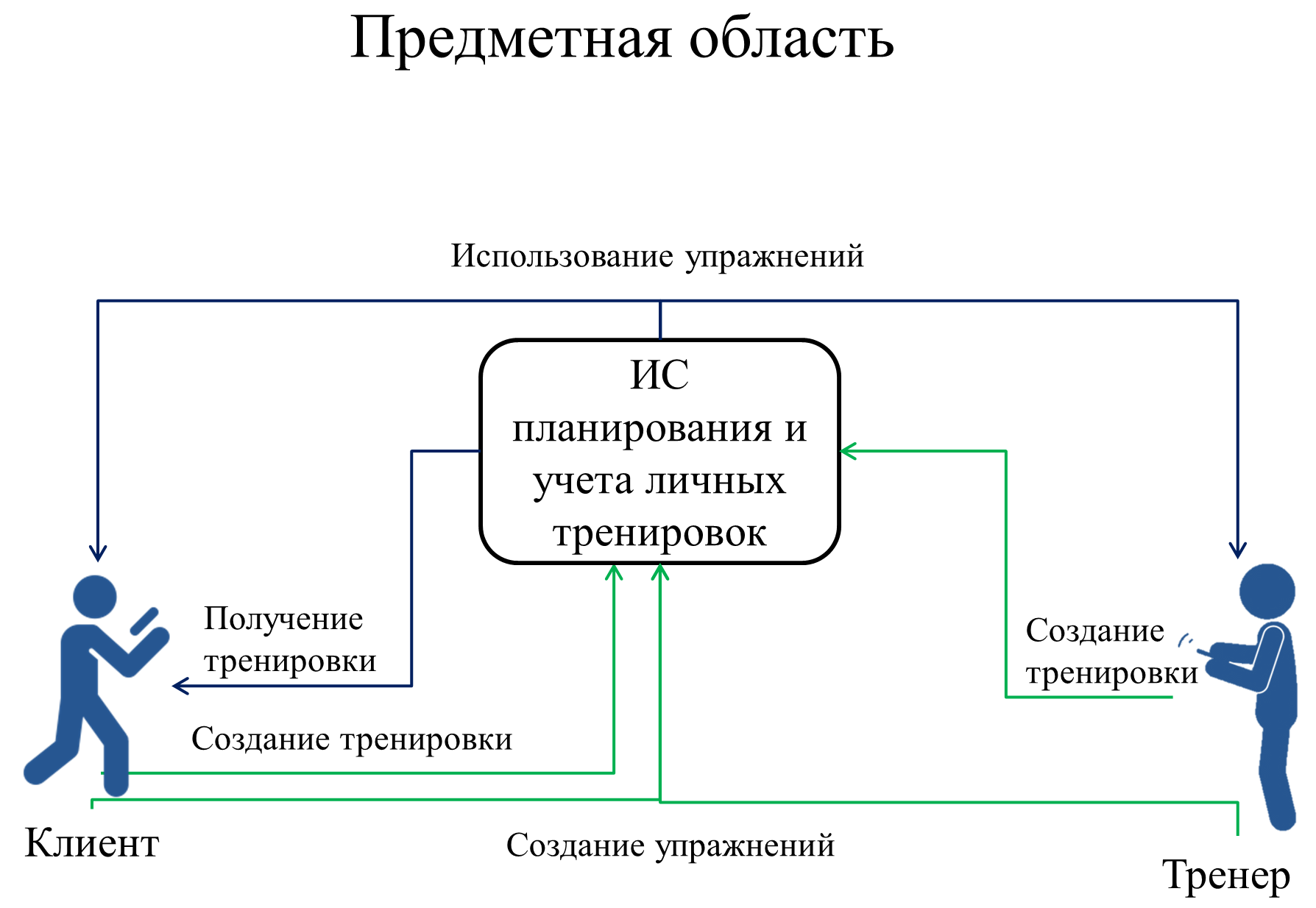
5. Инфологическая модель

6. Даталогическая модель

7. Пользовательские формы отображения

8. Пользовательские формы ввода





ПРИЛОЖЕНИЕ В   
ТЕХНИЧЕСКОЕ ЗАДАНИЕ

**Министерство образования и науки Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего профессионального образования**

**«Московский государственный технический университет имени Н.Э. Баумана»**

**(МГТУ им. Н.Э.Баумана)**

Утверждаю Согласовано

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_Черненький М.В.

«\_\_\_»\_\_\_\_\_\_\_\_\_\_2019 г. «\_\_\_»\_\_\_\_\_\_\_\_\_\_2019 г.

**«Информационная система планирования и учёта личных тренировок»**

техническое задание

(вид документа)

писчая бумага

(вид носителя)

9

(количество листов)

Исполнитель:

студент группы ИУ5-82

Фамилия И. О..

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_» \_\_\_\_\_\_\_\_\_\_2019 г.

Москва, 2019 г.

**СОДЕРЖАНИЕ**

[1 Наименование 3](#_Toc199783522)

[2 Основание для разработки 3](#_Toc199783523)

[3 Исполнитель 3](#_Toc199783524)

[4 Назначение и цель разработки 3](#_Toc199783525)

[5 Содержание работы 3](#_Toc199783526)

[5.1 Задачи, подлежащие решению 3](#_Toc199783527)

[5.2 Требования к функциональности программного изделия 4](#_Toc199783528)

[5.3 Требования к архитектуре программного изделия 5](#_Toc199783529)

[5.4 Требования к составу программных компонентов 5](#_Toc199783530)

[5.5 Требования к базе данных 5](#_Toc199783531)

[5.6 Требования к входным и выходным данным 5](#_Toc199783532)

[5.6.1 Требования к входным данным 5](#_Toc199783533)

[5.6.2 Требования к выходным данным 6](#_Toc199783534)

[5.7 Требования к составу и характеристикам программных средств 6](#_Toc199783535)

[5.7.1 Требования к программному обеспечению сервера 6](#_Toc199783536)

[5.7.2 Требования к программному обеспечению клиента 6](#_Toc199783537)

[5.8 Требования к составу и характеристикам технических средств 6](#_Toc199783538)

[5.8.1. Требования к составу и характеристикам технических средств сервера 6](#_Toc199783539)

[5.8.2 Требования к составу и характеристикам технических средств клиента 7](#_Toc199783540)

[6 Этапы разработки 7](#_Toc199783541)

[7 Техническая документация, предъявляемая по окончании работы 7](#_Toc199783542)

[8 Порядок приемки работы 8](#_Toc199783543)

9 [Дополнительные условия 8](#_Toc199783544)

**Наименование**

Информационная система планирования и учета личных тренировок. Краткое название: ИСПУЛТ

1. **Основание для разработки**

Основанием для разработки является задание на ВКР, подписанное руководителем ВКР и утвержденное заведующим кафедрой ИУ5 МГТУ им. Н.Э. Баумана.

1. **Исполнитель**

Студент МГТУ им. Н.Э. Баумана группы ИУ5-72 Горбовцова Ксения Михайловна.

1. **Назначение и цель разработки**

Разрабатываемая система предназначена для оптимизации общения тренеров с их клиентами при удаленной работе, а также для составления тренировочных планов индивидуального использования.

1. **Содержание работы**
   1. **Задачи, подлежащие решению**

* исследование предметной области;
* разработка программного изделия;
* отладка программного изделия;
* разработка инфологической модели базы данных;
* разработка даталогической модели базы данных;
* разработка графа диалога пользователя;
* исследование особенностей функционирования продукта в сети.
  1. **Требования к функциональности программного изделия**

ИСПУТ должна удовлетворять следующим требованиям:

Для пользователя системы должны быть обеспечены следующие возможности:

* + 1. Регистрация в системе;

1. Первичная регистрация;
2. Вход в систему;
3. Выход из системы;
   * 1. Создание собственных тренировок:
4. Добавление пользователей к тренировке;
5. Ввод данных тренировки
6. Добавление упражнений к тренировке;
   * 1. Создание упражнений:
7. Прикрепление оборудования к упражнению;
8. Ввод данных упражнения
9. Прикрепление параметра к упражнению;
   * 1. Работа с тренировками:
10. Просмотр тренировок с текущей датой;
11. Просмотр легенды тренировок;
12. Удаление тренировок;
13. Просмотр упражнений, входящих в тренировку;
    * 1. Работа с упражнениями:
14. Доступ к банку упражнений информационной системы;
15. Просмотр детализированной информации об упражнении;
16. Удаление упражнения из тренировки;
17. Поиск упражнения по названию в банке системы;
    * 1. Работа с другими зарегистрированными пользователями:
18. Поиск пользователя по email адресу;
19. Добавление пользователя в друзья;
20. Удаление пользователя из друзей;
21. Просмотр профиля пользователя;
    * 1. Работа с личным кабинетом:
22. Изменение фотографии;
23. Изменение пароля;
24. Изменение фамилии, имени и email адреса;

ПРИЛОЖЕНИЕ В   
ПРОГРАММА И МЕТОДИКА ИСПЫТАНИЙ

**Министерство образования и науки Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего профессионального образования**

**«Московский государственный технический университет имени Н.Э. Баумана»**

**(МГТУ им. Н.Э.Баумана)**

Утверждаю Согласовано

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_ Черненький М.В.

«\_\_\_»\_\_\_\_\_\_\_\_\_\_2019 г. «\_\_\_»\_\_\_\_\_\_\_\_\_\_2019 г.

**«Информационная система планирования и учёта личных тренировок»**

программа и методика испытаний

(вид документа)

писчая бумага

(вид носителя)

10

(количество листов)

Исполнитель:

студент группы ИУ5-82

Фамилия И.О..

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_»\_\_\_\_\_\_\_\_\_\_2019 г.

Москва ,2019 г.