**Введение**

Для обращения к системе связных данных (LOD) используется язык SPARQL. Но для людей не знакомых с ним это делает использование LOD абсолютно невозможным (SPARQL имеет высокий порог входа). Для решения этой проблемы исследователями, работающими над данной проблемой, разрабатываются естественно-языковые интерфейсы. Они позволяют обращаться к связным данным с помощью запросов на естественном языке. Но все работы, сделанные в данной области, в качестве естественного языка предполагают использование английского или его подъязыков. Поэтому, целью данной работы ставится разработка интеллектуального естественно-языкового интерфейса для обращения к LOD на русском языке.

В главе «Краткие сведения о семантической информационной системе LOD и ее применениях» будет рассказано о том, что такое система LOD, на каких принципах она основана, о том, что для представления данных в ней используется языковая система описания ресурсов RDF и приведены сведения об уже существующих проектах, в которых применяется система LOD.

В главе «Основные конструкции языка запросов SPARQL» рассказывается об языке запросов к сиcтеме LOD SPARQL, описывается его синтаксис и приводятся примеры основных конструкций языка и предлагаются примеры их применения.

Глава «Основные современные подходы к формальному описанию семантической структуры текстов на естественном языке (ЕЯ)» посвящена трем основным подходам к формальному описанию семантической структуры текстов: абстрактному представлению смысла (АПС), грамматике Монтегю и теория К-представлений (концептуальных представлений) В.А. Фомичева (Фомичев 2005). АПС и грамматика Монтегю специализированы на семантическом представлении английского языка, и с их помощью возможно описать только СП простых предложений. В свою очередь, теория К-представлений дает формальный аппарат для описания семантической структуры связных текстов (дискурсов), а не только отдельных предложений, на русском языке и других языках. Учитывая все эти факторы, для формального описания СП запроса на ЕЯ выбрана теория К-представлений.

Глава «Основные подходы к разработке семантически-ориентированных ЕЯ-интерфейсов для взаимодействия с прикладными интеллектуальными системами» рассмотрены подходы к разработке алгоритмов получения семантического представления текста на ЕЯ. Исследователи, использующие в своих работах семантическое представление, применяют последовательный лексический, синтаксический, семантический анализы с последующим построением семантического представления. В свою очередь в теории К-представлений В.А. Фомичев предлагает новый метод построения семантического представления текста на ЕЯ (Фомичев 2005). В рамках нового метода синтаксический и семантический анализ проводится одновременно и не используется синтаксический уровень представления текста. Упомянутый выше метод предполагает 3 этапа: компонентно-морфологический анализ входного текста, построение матричного семантико-синтаксического представления (МССП) и непосредственно сборка семантического представления текста по его МССП.

В главе «Основные подходы к разработке интеллектуальных интерфейсов для преобразования запроса к LOD на ЕЯ в запросы на языке SPARQL» рассматриваются подходы, применяемые в проектах, осуществляющих обращения к LOD на ЕЯ, для перевода запроса на ЕЯ в запрос на языке SPARQL. В частности, рассматриваются подходы, применяемые французским исследователем Себастиеном Ферре в своих работах «SQUALL: a Controlled Natural Language for Querying and Updating RDF Graphs» и «SQUALL: A Controlled Natural Language as Expressive as SPARQL 1.1» (Sébastien Ferré 2012, 2013), и американскими исследователями Нели Златаревой и Девишем Амином в работе «Natural Language to SPARQL Query Builder for Semantic Web Applications» (Neli Zlatareva, Devansh Amin 2021).

**Глава 1**

**Краткие сведения о семантической информационной системе LOD и ее применениях**

LOD (Linked Open Data) – открытые связные данные – система, подразумевающая использование методов публикации и связывание структурированных данных в Интернете. Основывается на четырех принципах, введенных Тимом Бернерс-Ли [1]:

1. Необходимо использование URI для объектов реального мира и абстрактных понятий, а не только для веб-документов и цифровых ресурсов;
2. Необходимо использование HTTP URI, что позволит использовать протокол HTTP для доступа к ресурсам;
3. Необходимо использование RDF (Resource Description Framework) в качестве единой модели данных для публикации данных;
4. Необходимо включать RDF утверждения на другие URI таким образом, чтобы была возможность обнаружения связанных данных.

В целом систему LOD можно представить, как огромный размеченный ориентированный граф, состоящий из элементарных графов, являющихся тройками (утверждениями) языка RDF.

RDF (Resource Description Framework)[2] – фреймворк описания ресурсов. Стоит сказать, что описание ресурсов (автор документа, дата публикации и т.п.) – это предполагаемое использование RDF. На практике он используется (и интерпретируется) как простейший язык для создания распределённых баз знаний. RDF представляет собой совокупность утверждений о ресурсах в виде, удобном для машинной обработки (формат RDF).

Формальная модель RDF состоит из следующих множеств:

* Множество «Ресурсы»;
* Множество «Литералы»;
* Подмножество «Ресурсов» – «Свойства»;
* Множество «Утверждения».

Последнее множество состоит из троек вида:

{субъект, предикат, объект},

где субъект – это элемент множества «Ресурсы»,

предикат – это элемент множества «Свойства»,

объект – это элемент либо множества «Ресурсы», либо множества «Литералы».

Множество «Ресурсы» состоит из веб-адресов информационных страниц, содержащих описания соответствующих ресурсов. Соответственно, множество «Свойства» – содержащих описания соответствующих предикатов. Множество «Литералы» состоит из всевозможных литералов, применение которых допустимо в описании тройки-утверждения.

Также существует языковая система RDFS [3], являющаяся надстройкой над языком RDF. Полное ее название RDF Schema Specification Language, но чаще всего используется сокращение RDF Schema (RDFS – рекурсивный акроним от RDF Schema). Система RDFS расширяет выразительные средства RDF следующими двумя возможностями: выделение и описание подклассов и указание семантических ограничений на атрибуты бинарных отношений, на аргументы и значения функций.

На основе опыта RDF и RDFS в 2004 создается язык проектирования онтологий (баз знаний) Ontology Web Language (OWL) [4]. На его основе на текущий момент разработано огромное количество онтологий по разным предметным областям.

OWL обладает двумя основными отличительными чертами, а именно: имеется много способов описания подклассов объектов и есть возможность указания свойств бинарных отношений (таких как, рефлективность, транзитивность, симметричность и т.п.) и использования этих свойств при семантическом поиске.

Существует большое количество задач, для решения которых может быть использована система взаимосвязанных данных LOD. В число таковых входят: моделирование и публикация пространственных данных, связывание и публикация данных из научно-исследовательских областей и бизнеса, встраивание данных научных исследований в LOD. Увеличение скорости поиска и анализа информации, в частности научно-исследовательской, медицинской, туристической и бизнес-информации.

Например, LOD используется в геоинформационном направлении. Геоинформационные данные нуждаются в логической и физической взаимосвязанности. Поэтому были организованы и проведены исследования, в ходе которых осуществлялся поиск способов преодоления семантической разнородности информации об географических объектах. В ходе данных исследований было введено понятие Инфраструктуры Пространственных Данных, были разработаны стандарты для построения систем на их основе.

Проекты LinkedGeoData (LGD) и GeoNames являются примерами реализации геоинформационных данных в системе LOD [5].

Проект LGD создавался для обеспечения преобразования и интеграции данных OSM (OpenStreetMap) в инфраструктуру системы LOD. А основным применением OSM является описание данных для визуализации различных карт. Другим проектом, использующим систему LOD, является географическая база данных GeoNames. В ней хранится около десяти миллионов имен географических объектов и состоит и семи с половиной тысяч уникальных элементов доступных для отображения на картах. GeoNames также интегрирует широкий спектр географических данных, такие как названия географических мест и объектов на разных языках, значения высот или глубин, разнообразные характеристики населения и другие, из разных источников.

В Малайзии для интеграции туристических ресурсов и связанности разрозненной информации о туризме была разработана семантическая рекомендательная система по туризму в Малайзии (SMTPS) [6]. Данная система позволяет найти отель, ресторан в заданном месте, узнать об активностях, проводимых в конкретный день, а также просто получить список достопримечательностей. При этом информация, используемая для ответа на запрос, может располагаться на разных информационных ресурсах.

Использование системы LOD в бизнесе позволяет предприятием быстрее и проще получать доступ к интересующим законодательным актам, действующим на территории страны, передавать избранный контент и информацию на аутсорсинг, для поддержки специалистами соответствующей предметной области [7].

Большое количество научной информации хранится в книгах, а книги в свою очередь в библиотеках. При всем своем удобстве библиотеки представляют собой огромный массив данных, в котором отдельно взятому человеку трудно ориентироваться, не говоря уже о поиске литературы по конкретной тематике. Для того чтобы пользователи библиотек имели быстрый и максимально возможный полный доступ, была разработана Классификационная система [8]. Классификационная система обеспечивает семантическую интеграцию библиотечных данных, посредством публикации последних в виде LOD. Классификационная система работает на основе данных Российской государственной библиотеки.

**Основные конструкции языка запросов SPARQL**

SPARQL (рекурсивный акроним SPARQL Protocol and RDF Query Language) – язык запросов к данным, представленным в формате RDF. SPARQL для LOD является таким же инструментом, как и SQL для реляционных баз данных. SPARQL имеет сходную с SQL форму запроса, а именно «SELECT FROM WHERE», т.е. «Что выбрать», «В каком ресурсе» и «Каким параметрам должно соответствовать» [9].

Простейший запрос SPARQL представлен в листинге 2. В качестве данных, к которым применяется запрос, используется RDF граф, приведенный в листинге 1.

Листинг 1 – RDF-тройка

<http://example.org/book/book1> <http://purl.org/dc/elements/1.1/title> "SPARQL Tutorial" .

Листинг 2 – Простейший запрос на языке SPARQL

SELECT ?title WHERE {<http://example.org/book/book1> <http://purl.org/dc/elements/1.1/title> ?title . }

Результатом выполнения данного запроса является строка «SPARQL Tutorial».

Запрос приведенный выше осуществляет поиск книги по заданному графу RDF. Запрос состоит из двух частей: части SELECT, в которой определяются переменные, которые будут отображаться в результатах запроса; и части WHERE, в которой представлен шаблон графа для сопоставления с графом данных, в данном запросе шаблон состоит из графа с одной переменной.

RDF предоставляет возможность использования литералов трех варианты литералов трех вариантов: с пометкой (тегом) языка, литерал со стандартным типом данных и литерал с произвольным типом данных (возможен еще с четвертый тип литерала – простой литерал без тегов языка и типов, но в следующих трёх примерах запросов SPARQL он рассматриваться не будет). Пример данных с использование трёх типов литералов приведены в листинге 3.

Листинг 3 – RDF граф с использованием литералов

@prefix dt: <http://example.org/datatype#> .

@prefix ns: <http://example.org/ns#> .

@prefix : <http://example.org/ns#> .

@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

:x ns:p "cat"@en .

:y ns:p "42"^^xsd:integer .

:z ns:p "abc"^^dt:specialDatatype .

В листингах 4 и 5 приведены на первый взгляд похожие запросы, но запрос в листинге 4 не будет иметь соответствие в указанных выше данных, в отличие от запроса в листинге 5, поскольку литералы «cat» и «cat@en» не являются одинаковыми.

Листинг 4 – Запрос с литералом без тега языка

SELECT ?v WHERE { ?v ?p "cat" }

Листинг 5 – Запрос с некорректным литералом

SELECT ?v WHERE { ?v ?p "cat@en" }

В листинге 6 приведен пример запроса для поиска литерала со стандартным типом. При использование в литерале стандартного типа возможно использование сокращенной формы записи, например просто 23 вместо "23"^^<http://www.w3.org/2001/XMLSchema#integer>. В этом случае задача определения конкретного типа возлагается на программное обеспечение реализующее выполнение запроса.

Листинг 6 – Пример запроса с литералом со стандартным типом

SELECT ?v WHERE { ?v ?p 42 }

Следующий пример — это запрос с литералом с произвольным типом (листинг 7). В данном случае литерал должен указываться полностью в явном виде, но при выполнении такого запроса обработчик запроса не обязан знать о указанном типе данных, поиск производится по сравнению литерала.

Листинг 7 – Пример запроса с литералом с произвольным типом

SELECT ?v WHERE { ?v ?p "abc"^^<http://example.org/datatype#specialDatatype> }

Результат выполнения запроса «http://example.org/ns#x».

Возможности ограничения выборки искомых ресурсов будет показано на основе данных, указанных в листинге 8.

Листинг 8 – Данные для демонстрации запросов с ограничениями на искомую выборку

@prefix dc: <http://purl.org/dc/elements/1.1/> .

@prefix : <http://example.org/book/> .

@prefix ns: <http://example.org/ns#> .

:book1 dc:title "SPARQL Tutorial" .

:book1 ns:price 42 .

:book2 dc:title "The Semantic Web" .

:book2 ns:price 23 .

Язык SPARQL позволяет ограничивать выборку по содержанию строковыми литералами определенных значений, по числовым литералам на основе арифметических признаков (больше, меньше и т.п.) также возможны органичения по логическому типу и по дате (до указанной даты, после и т.п.).

В листингах 9 и 10 приведены запросы, ограничивающий выборку, по содержимому строкового литерала.

Листинг 9 – Запрос, ограничивающий выборку, ресурсами, название которых начинается с подстроки «SPARQL»

PREFIX dc: <http://purl.org/dc/elements/1.1/>

SELECT ?title

WHERE { ?x dc:title ?title

FILTER regex(?title, "^SPARQL")

}

При применении запроса к данным из листинга 8, результат выполнения запроса будет следующим

|  |
| --- |
| **title** |
| "SPARQL Tutorial" |

Листинг 10 – Запрос, ограничивающий выборку, ресурсами, название которых содержит подстроку «web»

PREFIX dc: <http://purl.org/dc/elements/1.1/>

SELECT ?title

WHERE { ?x dc:title ?title

FILTER regex(?title, "web", "i" )

}

При применении запроса к данным из листинга 8, результат выполнения запроса будет следующим

|  |
| --- |
| **title** |
| "The Semantic Web" |

В листингах 11 и 12 приведены примеры запросов ограничивающих выборку на основе значений численных литералов.

Листинг 11 – Запрос с ограничением выборки по числовому литералу

PREFIX dc: <http://purl.org/dc/elements/1.1/>

PREFIX ns: <http://example.org/ns#>

SELECT ?title ?price

WHERE { ?x ns:price ?price .

FILTER (?price < 30.5)

?x dc:title ?title . }

При применении запроса к данным из листинга 8, результат выполнения запроса будет следующим

|  |  |
| --- | --- |
| **title** | **price** |
| "The Semantic Web" | 23 |

Запросы с ограничением по логическому типу и дате имеют структуру запроса аналогичную запросам с ограничением по числовому типу.

**Основные современные подходы к формальному описанию семантической структуры текстов на естественном языке (ЕЯ)**

Одним из подходов к формальному описанию текстов на ЕЯ является АПС (Абстрактное представление смысла) [10,11] (Banarescu, L. et al 2013, 2019). В рамках данного подхода каждое предложение представляется в виде корневого, направленного ациклического графа с метками на ребрах (отношения между понятиями) и листьях (сами понятия). В качестве понятий в АПС используются английские слова («boy» как в примере выше), фреймы из PropBank («believe-01») или ключевые слова, включающие специальные типы сущностей: даты, региона, расстояния и др, и союзы связки: «и», «или».

АПС учитывает все слова предложения, но при этом он является абстрактным, то есть один граф может представлять несколько разных предложений на естественном языке со схожим смыслом. В качестве примера можно привести следующий набор предложений:

* The boy desires the girl to believe him;
* The boy desires to be believed by the girl;
* The boy has a desire to be believed by the girl;
* The boy’s desire is for the girl to believe him.

У перечисленных выше предложений будет одинаковое описание семантической структуры (листинг 1).

Листинг 1 – Пример АПС

(w / want-01

:ARG0 (b / boy)

:ARG1 (b2 / believe-01

:ARG0 (g / girl)

:ARG1 b))

В этом примере проявляется возможность АПС абстрагироваться от синтаксической структуры, что позволяет выделять предложения с разной синтаксической структурой, но с одинаковой семантической.

Хоть АПС и не зависит от синтаксических особенностей, но он рассчитан для описания семантической структуры предложений только на английском языке и без взаимосвязи с предыдущими предложениями.

Также существует подход к формальному описанию семантической структуры текста называемый Грамматика Монтегю [12]. В рамках данного подхода семантическая структура описывается с помощью формальных инструментов логики предикатов первого порядка и лямбда-исчисления. При разработке своей грамматики автор руководствовался собственным утверждением о том, что естественные и формальные языки в существенных своих свойствах не отличаются друг от друга, что в свою очередь позволяет описывать их семантику и синтаксис в рамках одной математической модели. В качестве примера можно привести СП предложения «there is man»:

S → there is NP { np λx.true } “there is [NP a man]”.

Данному подходу свойственны недостатки аналогичные недостаткам АПС, а именно то, что Грамматика Монтегю рассчитана на описание предложений на английском языке, она не располагает формальными инструментами для описания дискурсов. Также стоит отметить, что грамматика изложена автором в трудно воспринимаемой форме и является трудной для реализации в программном виде из-за вычислительной сложности.

Выше указанным подходам свойствена изначальная ориентация на описания высказываний (в Грамматике Монтегю еще есть возможность описывать вопросы, но этого все равно недостаточно). А как бы то не было, еще существуют выражения, соответствующие целям, действиям, определениям, назначениям вещей, поэтому должен быть и формальный аппарат, позволяющий описывать смысловую структуру таких выражений.

Последний рассматриваемый и применяемый в данной работе подход для формального описания семантической структуры текстов на ЕЯ –теория К-представлений (концептуальных представлений) В.А. Фомичева [13] (Фомичев 2005). Данная теория описывает математическую модель, позволяющую формально описать семантическое представление текстов (дискурсов) на ЕЯ. При этом данный подход не обладает недостатками АПС и грамматики Монтегю, которые могут описать СП предложения предложений на подъязыках английского. С помощью теории К-представлений возможно описать СП текстов в первую очередь на русском, а также на английском, немецком и французском языках. Также в рамках данного подхода возможно, строить представления как фраз, выражающих высказывания, так и повествовательных текстов; представления целей (выраженных неопределенными формами глаголов с зависимыми словами) и вопросов; строить и различать обозначения единиц, соответствующих объектам, ситуациям, процессам в реальном мире и понятиям, характеризующим эти объекты, ситуации и процессы; строить и различать обозначения: объектов и множеств объектов, понятий и множеств понятий.

В качестве примера К-представления можно привести семантическое представление следующего предложения: «Сколько раз в этом году запрашивался учебник Коробова?». Для данного будет построено следующее СП:

Вопрос (x1, (x1 ≡ Колич-элем(все запрос1 \*(Время, текущий-год) (Предмет-запроса, нек учебник \* (Автор, нек человек \*(Фамилия, “Коробов”) : x2))))) .

**Основные подходы к разработке семантически-ориентированных ЕЯ-интерфейсов для взаимодействия с прикладными интеллектуальными системами**

В работах зарубежных исследователей (B. Nethravathi et al 2020, Lee M. Christensen et al 2009, Chuan Wang et al 2015, Sylvain Pogodalla 2004) [14-17] в основном применяется совпадающий в значимых частях подход для построения семантического представления текста на ЕЯ. Он в общем виде состоит из последовательно выполняемых лексического, синтаксического, семантического анализов и непосредственного построения представления.

В ходе лексического анализа осуществляется деление текста на лексические единицы и определяются их морфологические признаки. Определение признаков проводится как с помощью запросов к разнообразным лингвистическим банкам с соответствующей информацией, так и с помощью средств лексического анализа собственной разработки.

На следующем этапе, синтаксического анализа, определяются взаимосвязи между лексическими единицами и строится синтаксическое дерево.

Далее на основе данных полученных на предыдущих этапах проводится семантический анализ, в ходе которого определяется смысловая структура входного текста.

И на основе результатов семантического анализа проводится непосредственно построение семантического представления текста.

В свою очередь Владимир Александрович Фомичев в своей научной монографии «Формализация проектирования лингвистических процессоров» в рамках своей теории К-представлений формулирует новый метод преобразования ЕЯ-текста в СП текста [13] (Фомичев 2005). Данный метод предназначается для разработки ЕЯ-интерфейсов для взаимодействия с прикладными интеллектуальными системами. Указанный выше метод состоит из трех частей (этапов) преобразования.

На первом этапе осуществляется компонентно-морфологический анализ входного текста. По входному тексту строится не менее одного компонентно-морфологических представления (КМП), каждое из которых представляет из себя пару (набор) классифицирующего и морфологического представлений текста. Последнее представляет возможные значений морфологических признаков для лексических единиц входной текста. Чаще всего отдельным фразам соответствует единственное КМП. В ситуации неоднозначного разбиения на значащие единицы или невозможности однозначного определения части речи какой-либо единицы пользователю задаются уточняющие вопросы, ответы на которые позволяют снять неоднозначность.

Второй этап предполагает построение матричного семантико-синтаксического представления (МССП). На этом этапе с каждым словом связывается одно из возможных значений и устанавливаются смысловые отношения между единицами текста.

МССП сначала недоопределенно и неопределенность снимется шаг за шагом. Для снятия неопределенности используется информация из лингвистической базы данных (ЛДБ) о возможных способах соединения единиц текста в лингвистически допустимые сочетания.

На заключительном, третьем этапе производится сборка семантического представления текста, которое является его   
К-представлением, по его МССП.

С помощью описанного выше метода возможно устанавливать смысловые соотношения в сочетаниях типа: «Глагол + Предлог + Существительное», «Глагол + Существительное», «Существительное1 + Предлог + Существительное2», «Число + Существительное», «Прилагательное + Существительное», «Существительное1 + Существительное2», «Причастие + Существительное», «Причастие + Предлог + Существительное», «Вопросительно-относительное местоимение или местоименное наречие + Глагол», «Предлог + Вопросительно-относительное местоимение + Глагол».

Довольно важно отметить, что данный метод учитывает многозначность слов, что чрезвычайно важно при проектировании ЕЯ-интерфейсов для взаимодействия с прикладными интеллектуальными системами. А также не предполагает использование синтаксического уровня представления текста (как следствие и синтаксического анализа), хоть синтаксический уровень и используется в течение долго времени как отечественными, так и зарубежными исследователями.

**Основные подходы к разработке интеллектуальных интерфейсов для преобразования запроса к LOD на ЕЯ в запросы на языке SPARQL**

Аналогичными работа по разработке интеллектуальных интерфейсов для преобразования запроса к LOD на ЕЯ в запросы на языке SPARQL занимается французский исследователь Себастьен Ферре (Sébastien Ferré 2012, 2013). Одной из его разработок является SQUALL. Он описан в двух работах Ферре: «SQUALL: a Controlled Natural Language for Querying and Updating RDF Graphs» [18] и «SQUALL: A Controlled Natural Language as Expressive as SPARQL 1.1.». [19]

SQUALL является контролируемым языком, т.е. такой версией естественного языка, которая получена с помощью ограничения в использовании грамматической вариативности, определенных речевых оборотов. Перечисленные выше ограничения проводятся с целью устранения (уменьшения) многозначности и сложность, что в свою очередь обеспечивает формальную логическую основу, т.е. формальные семантику и синтаксис. А язык с формальными семантикой и синтаксисом может быть сопоставлен с другим формальным языком. При разработке языка SQUALL Ферре придерживался следующих двух требований:

1. язык должен обладать выразительностью сопоставимой с выразительностью языка SPARQL;
2. обладая выразительностью SPARQL, язык должен быть высокоуровневым и иметь естественный синтаксис, абстрагированный от низкоуровневых понятий, таких как реляционная алгебра, которая используется в языке SPARQL.

Иными словами, французский исследователь разработал по факту естественно-языковой аналог SPARQL.

Для перевода из языка SQUALL в язык SPARQL используется промежуточное представление – грамматика Монтегю. Для упрощения представления и дальнейшего перевода в SPARQL некоторые конструкторы заменяются на понятия, которые они отображают.

Перевод из промежуточного языка в SPARQL осуществляется с помощью сравнения логических конструкций с конструкциями языка SPARQL.

Формулы с конструкторами переводятся в запросы SPARQL, другие типы формул переводятся в запросы обновления SPARQL. Встреча при переводе переменной предполагает создание новой переменной SPARQL. Предикаты преобразуются в фильтры SPARQL, агрегации – в агрегативные подзапросы SPARQL.

В работе «Natural Language to SPARQL Query Builder for Semantic Web Applications» [20] (Neli Zlatareva, Devansh Amin 2021) применяется подход отличный от подхода Ферре. В данной работе не разрабатывался специальный язык, а уклон сделан на выделении необходимых для запроса на языке SPARQL смысловых структур из синтаксического дерева зависимостей.

Первым этапом является распознавание именованных сущностей, содержащихся в запросе на ЕЯ. Далее строится синтаксическое дерево зависимостей, на основе которого далее определяются предикаты, которые будут использоваться при построении запроса на языке SPARQL. Следующим шагом применяется классификатор типа запроса для определения типа запроса на ЕЯ, что далее повлияет на структуру запроса на языке SPARQL (возможны три типа запроса: единичный факт, единичный факт с конкретным типом и вопрос). На следующем шаге производится сопоставление предикатов и сущностей, полученных на предыдущих шагах, с предикатами и сущностями, определёнными в онтологии приложения. Также на этом этапе преодолеваются лексические пробелы и смысловые разрывы (появляются из-за содержательной разницы или разницы между представлением онтологии и информацией, требуемой в запросе). Для этого применяется алгоритм, обеспечивающий сравнение на основе косинусного сходства. На заключительном этапе на основе предикатов и сущностей, предоставленных на предыдущем этапе, строится запрос SPARQL.

**Выводы**

В данной главе сделан обзор актуальности и сфер применения системы Открытых Связанных Данных (LOD). Использование LOD позволяет получать требуемую информацию быстрее и точнее, чем при использовании классических способов поиска информации.

Рассмотрены основные конструкции языка SPARQL.

Проанализированы основные подходы для формального представления семантической структуры текста на естественном языке. Наиболее подходящим подходом для использования в данной работе является теория К-представлений В.А. Фомичева, поскольку она предоставляет формальный аппарат для описания семантической структуры русского языка.

Проведен анализ подходов к разработке семантически-ориентированных ЕЯ-интерфейсов для взаимодействия с прикладными интеллектуальными системами. Для использования в данной работе был выбран новый метод преобразования текста в его семантическое представление предложенный В.А. Фомичевым.

Рассмотрены подходы к переводу запроса на естественном языке в запрос на языке SPARQL. Ни один из имеющихся подходов не обеспечивает использование запросов на русском языке.

**Глава 2**

**Разработка логической структуры   
лингвистической базы данных**

Для преобразования текста на ЕЯ в его семантическое представление с помощью нового метода преобразования, предложенного В.А. Фомичевым (Формализация проектирования лингвистических процессоров, 2005) предполагается использование промежуточного представления – матричное семантико-синтаксическое представление (МССП) [13]. В начальный момент времени МССП недоопределенно и неопределенность снимается последовательно, шаг за шагом на основе информации из лингвистической базы данных (ЛБД) о возможных способах соединения единиц текста в лингвистически корректные сочетания. Для этого в ЛБД должна содержать информацию, позволяющую устанавливать возможные семантические отношения следующих сочетаний: «Глагол + Предлог + Существительное», «Глагол + Существительное», «Существительное1 + Предлог + Существительное2», «Число + Существительное», «Прилагательное + Существительное», «Существительное1 + Существительное2», «Причастие + Существительное», «Причастие + Предлог + Существительное», «Вопросительно-относительное местоимение или местоименное наречие + Глагол», «Предлог + Вопросительно-относительное местоимение + Глагол».

Для выполнения поставленных условий ЛБД должна состоять из 4 компонентов: морфологической базы данных, лексико-семантического словаря, словаря глагольно-предложных семантико-синтаксических фреймов и словаря предложных семантико-синтаксических фреймов [13].

Морфологическая база данных хранит информацию о конкретной словоформе в виде набора значений, следующих параметров: части речи, подкласса части речи, падежа, числа, рода, залога, времени, наклонения, вида, лица, возвратности. Для одной словоформы возможно наличие нескольких наборов. Например, со словом "книги" может быть связано три набора значений морфологических признаков (если "книги" - словоформа в единственном числе, то эта словоформа находится в родительном падеже; если "книги" - словоформа во множественном числе, то она может быть, как в именительном, так и в винительном падежах).

Лексико-семантический словарь является одним из основных компонентов ЛБД, ставящим в соответствие единицам текстов единицы семантического уровня. В общем виде его структуру можно представить следующим образом: (Номер набора, Лексема, Часть речи лексемы, Сем, Семантические координаты, Комментарий).

Номер набора служит для однозначной идентификации набора, что позволит обеспечить организацию циклов в алгоритме преобразования текста в семантическое представление.

Лексема – элемент множества лексем из морфологической базы данных.

Сем – соответствующая лексеме семантическая единица. Для глаголов, причастий, деепричастий семантическая единица, связанная с соответствующим отглагольным существительным. Например, глагол "поступить" имеет два значения: поступление абитуриента в учебное заведение; поступление физического объекта на какой-то пространственный объект (например, товара на склад). Поэтому началом одного из наборов возможного лексико-семантическим словаря будет последовательность элементов (n1, поступить, глагол, поступление1), а началом другого набора - последовательность (n2, поступить, глагол, поступление2). Для прилагательного Сем представляется в виде предиката. Например, для лексемы «зеленый» семом будет (Цвет(z1, зеленый)).

Семантические координаты сущности, характеризуемой Семом, определяют различную смысловую нагрузку, которая может быть у нее, в разных контекстах. Например, если Сем – фирма, то возможными семантическими координатами может быть набор (Интеллектуальная система, Пространственный объект, Организация).

Лексико-семантический словарь может включать в себя следующие наборы:

(208, поступить, глаг, поступление1, сит, «поступить в вуз»),

(209, поступить, глаг, поступление2, сит, nil, nil, «поступил груз»),

(311, алюминевый, прилаг, Материал(z1, алюминий),физ.об),

(358, зеленый, прилаг, Цвет(z1, зелен), физ.об),

(411, пассажирский, прилаг, Назначение(z1, перемещение1 \* (Объект1, опред множ \* (Кач-состав, человек))), дин.физ.об).

В формировании предложения определяющую роль играют глаголы, причастия, деепричастия и отглагольные существительные. Они выражают отношения между объектами, участвующими в предложении.

Смысловое отношение между значением глагольной формы (личной формы глагола, неопределенной формы глагола, причастия, деепричастия, отглагольного существительного) и значением зависящей от нее группы слов называется тематической ролью (семантическим падежом). В разрабатываемой ЛБД тематические роли описываются в словаре глагольно-предложных фреймов.

Словарь глагольно-предложных фреймов состоит из набора со следующей структурой: (Номер набора, Семантическая единица, описывающая ситуацию, Форма глагола, Возвратность глагола (причастия), Залог (страдательный или действительный), Предлог (возможен и составной), Падеж зависимого существительного, Семантическое ограничение значения зависимой группы слов, Тематическая роль, Поясняющий пример (необязателен)).

Пример возможных наборов, содержащихся в словаре глагольно-предложных фреймов:

(1, подготовка1, личн, нвз, действ, nil, 1, интс, Агент1, ‘И.П.Сомов

подготовил (учебное пособие) ’),

(2, подготовка1, личн, нвз, страд, nil, 5, интс, Агент1, 'Профессором Семеновым была подготовлена (книга)’),

(3, подготовка1, личн, нвз, действ, nil, 4, инф.об, Продукт1, ‘(И.П.Сомов) подготовил книгу’ ),

(4, подготовка1, личн, нвз, страд, nil, 1, инф.об, 1, инф.об, ‘Статья была

подготовлена (за три недели) ’).

В словаре предложных семантико-синтаксических фреймов содержаться наборы, описывающие возможные семантические отношения либо между двумя связанными существительными, либо между двумя существительными, связанными предлогом.

Словарь предложных семантико-синтаксических фреймов состоит из наборов следующего вида: (Номер набора, предлог (в том числе нулевой), Возможная семантическая координата первого существительного, Возможная семантическая координата второго существительного, Падеж второго существительного, Обозначение смыслового отношения, Пример отношения (необязателен)).

В качестве примера словаря семантико-синтаксических фреймов можно привести следующие наборы (фреймы):

(1, ‘от’, вещество, болезнь, 2, Против1, ‘таблетки от гриппа’);

(2, ‘от’, вещество, дин.физ.об, 2, Против2, ‘мазь от комаров’);

(3, ‘от’, физическое явление, физ.об, 2, Эффект1, ‘тень от дома’).

**Описание разработанной схемы лингвистической базы данных**

На основе описанной выше логической структуры была разработана схема лингвистической базы данных. Ее описание по компонентам предоставлено дальше.

*Морфологическая база данных*

Для морфологической базы данных, схема представлена на рисунке 1.

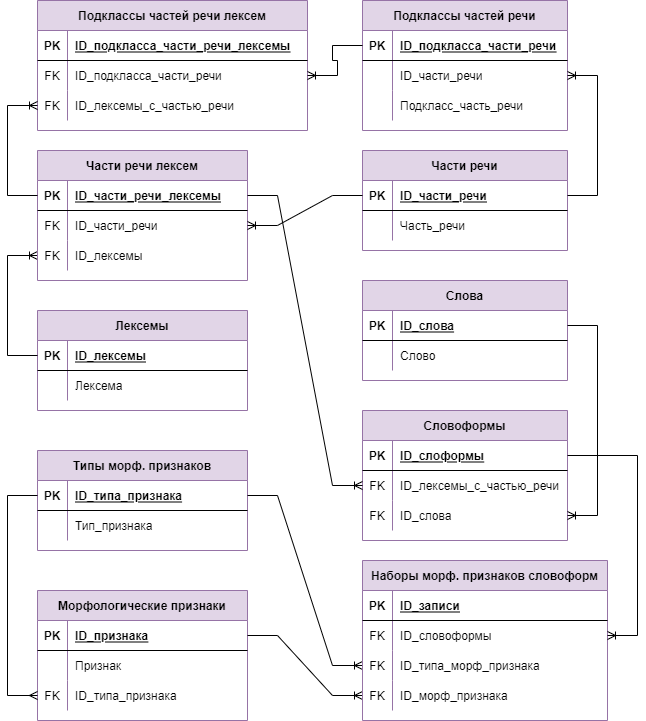


Рисунок 1 – Схема морфологической базы данных

*Описание таблиц МБД*

**Таблица «Слова»**

Содержит слова естественного языка (уникальные)

*Поля таблицы «Слова»:*

* ID\_слова – идентификатор конкретного слова (Тип данных: целое число);
* Слово – конкретное уникальное слово (Тип данных: строка переменного размера).

**Таблица «Части речи»**

Содержит список частей речи.

*Поля таблицы «Части речи»:*

* ID\_части\_речи – идентификатор конкретной части речи (Тип данных: целое число);
* Часть\_речи – название части речи (Тип данных: строка переменного размера).

**Таблица «Подклассы частей речи»**

Содержит список подклассов частей речи.

*Поля таблицы «Части речи»:*

* ID\_подкласса\_части\_речи – идентификатор конкретного подкласса части речи (Тип данных: целое число);
* Подкласс\_часть\_речи – название подкласса части речи (Тип данных: строка переменного размера);
* ID\_части\_речи – идентификатор части речи (Тип данных: целое число).

**Таблица «Лексемы»**

Содержит список лексем (уникальных).

*Поля таблицы «Лексемы»:*

* ID\_лексемы – идентификатор конкретной части речи (Тип данных: целое число);
* Лексема – конкретная уникальная лексема (Тип данных: строка переменного размера).

**Таблица «Части речи лексем»**

Таблица связи лексем с соответствующими им частями речи.

*Поля таблицы «Части речи лексем»:*

* ID\_части\_речи\_лексемы – идентификатор записи о связи лексемы с ее частью речи (Тип данных: целое число);
* ID\_части\_речи – идентификатор конкретной части речи (Тип данных: целое число);
* ID\_лексемы – идентификатор конкретной лексемы (Тип данных: целое число).

**Таблица «Подклассы частей речи лексем»**

Таблица связи лексем с соответствующими им подклассами частямей речи.

*Поля таблицы «Части речи лексем»:*

* ID\_подкласса\_части\_речи\_лексемы – идентификатор записи о связи лексемы с ее поклассом частью речи (Тип данных: целое число);
* ID\_подкласса\_части\_речи – идентификатор подкласса части речи (Тип данных: целое число);
* ID\_лексемы – идентификатор строки таблицы «Части речи лексемы» (Тип данных: целое число).

**Таблица «Словоформы»**

Таблица связи слов с соответствующими им лексемами.

*Поля таблицы «Словоформы»:*

* ID\_словоформы – идентификатор конкретной словоформы (Тип данных: целое число);
* ID\_лексемы\_с\_частью\_речи – идентификатор записи таблицы «Части речи лексем», определяющей лексему с частью речи (Тип данных: целое число);
* ID\_слова – идентификатор слова (Тип данных: целое число).

**Таблица «Типы морфологических признаков»**

Содержит список типов морфологических признаков (падеж, склонение и т.п.).

*Поля таблицы «Типы морфологических признаков»:*

* ID\_типа\_признака – идентификатор конкретной словоформы (Тип данных: целое число);
* Тип\_признака – название типа признака (Тип данных: строка переменного размера).

**Таблица «Морфологические признаки»**

Содержит список возможных значений морфологических признаков (родительный, 1 склонение и т.п.).

*Поля таблицы «Морфологические признаки»:*

* ID\_признака – идентификатор значения признака (Тип данных: целое число);
* Признак – значение признака (Тип данных: строка переменной длины);
* ID\_типа\_признака – идентификатор типа признака (Тип данных: целое число).

**Таблица «Наборы морфологических признаков словоформ»**

Содержит набор морфологических признаков словоформ.

*Поля таблицы «Наборы морфологических признаков словоформ»:*

* ID\_записи – идентификатор записи (Тип данных: целое число);
* ID\_словоформы – идентификатор словоформы (Тип данных: целое число);
* ID\_типа\_морф\_признака – идентификатор типа морфологического признака (Тип данных: целое число);
* ID\_морф\_признака – идентификатор значения морфологического признака (Тип данных: целое число).

*Лексико-семантический словарь*

Схема базы данных для лексико-семантического словаря представлена на рисунке 2.



Рисунок 2 – Схема лексико-семантического словаря

*Описание таблиц лексико-семантического словаря*

Таблица «Подклассы частей речи лексем» является частью МБД и описана в предыдущем разделе.

**Таблица «Типы семантических значений»**

Содержит набор возможных типов семантических значений.

*Поля таблицы «Типы семантических значений»:*

* ID\_типа\_значения – идентификатор типа семантического значения (Тип данных: целое число);
* Тип\_значения – название типа семантического значения (Тип данных: строка переменной длины).

**Таблица «Семантические значения»**

Содержит набор возможных семантических значений разных типов (основное значение, дополнительное значение, тематическая роль и т.п.).

*Поля таблицы «Семантические значения»:*

* ID\_значения – идентификатор семантического значения (Тип данных: целое число);
* ID\_типа\_значения – идентификатор типа семантического значения (Тип данных: целое число).

**Таблица «Основные семантические значения лексем»**

Содержит связь лексем с их основными семантическими значениями.

*Поля таблицы «Основные семантические значения лексем»:*

* ID\_основного\_значения\_лексемы – идентификатор значения конкретной лексемы (Тип данных: целое число);
* ID\_подкласса\_части\_речи\_лексемы – идентификатор конкретной лексемы с определенным подклассом части речи (Тип данных: целое число);
* ID\_основного\_значения – идентификатор основного семантического значения (Тип данных: целое число);
* Комментарий – необязательное поле, в котором может содержаться пример использования данной лексемы с заданным семантическим значением.

**Таблица «Дополнительные семантические значения лексем»**

Содержит связь лексем с их дополнительными семантическими значениями.

*Поля таблицы «Дополнительные семантические значения лексем»:*

* ID\_доп\_значения\_лексемы – идентификатор дополнительного значения конкретной лексемы (Тип данных: целое число);
* ID\_основного\_значения\_лексемы – идентификатор лексемы с определенным основным семантическим значением (Тип данных: целое число);
* ID\_дополнительного\_значения – идентификатор дополнительного семантического значения (Тип данных: целое число)*;*
* Комментарий – необязательное поле, в котором может содержаться пример использования данной лексемы с заданным семантическим значением.

*Словарь глагольно-предложных фреймов*

Схема базы данных для словаря предложных фреймов представлена на рисунке 3.

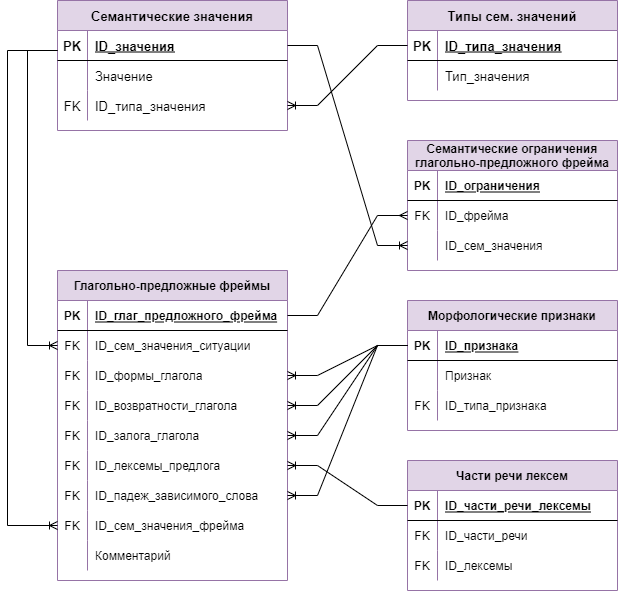


Рисунок 3 – схема словаря глагольно-предложных фреймов

*Описание таблиц словаря глагольно-предложных фреймов*

Таблицы «Части речи лексем» и «Морфологические признаки» описаны в рамках морфологической базы данных. Таблицы «Семантические значения» и «Типы семантических значений» описаны в рамках лексико-семантического словаря.

**Таблица «Глагольно-предложные фреймы»**

Содержит описания глагольно-предложных фреймов.

*Поля таблицы «Глагольно-предложные фреймы»:*

* ID\_глаг\_предложного\_фрейма – идентификатор глагольно-предложного фрейма (Тип данных: целое число);
* ID\_сем\_значения\_ситуации – идентификатор семантического значения ситуации, описываемой глагольной формой (Тип данных: целое число);
* ID\_формы\_глагола – идентификатор признака, определяющего форму глагола (Тип данных: целое число);
* ID\_возвратности\_глагола – идентификатор признака, определяющего возвратность глагола (Тип данных: целое число);
* ID\_залога\_глагола – идентификатор признака, определяющего залог глагола (Тип данных: целое число);
* ID\_лексемы предлога – идентификатор лексемы предлога, связанного с глагольной формой (Тип данных: целое число);
* ID\_падеж\_зависимого\_слова – идентификатор признака, определяющего падеж слова (Тип данных: целое число);
* ID\_сем\_значения\_фрейма – идентификатор семантического значения фрейма (Тип данных: целое число);
* Комментарий – необязательное поле, в котором может содержаться пример использования данного фрейма.

*Словарь предложных фреймов*

Схема базы данных для словаря предложных фреймов представлена на рисунке 4.

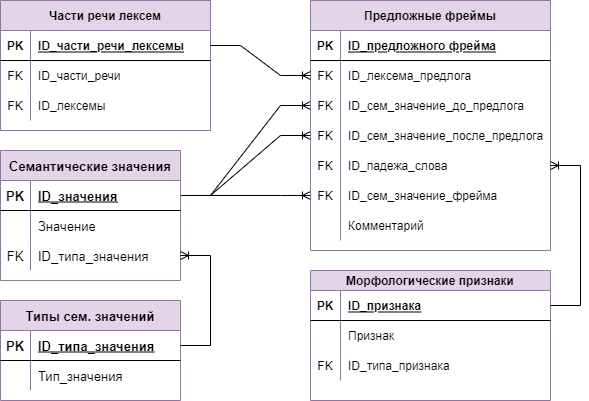


Рисунок 4 – схема словаря предложных фреймов

*Описание таблиц словаря предложных фреймов*

Таблицы «Части речи лексем» и «Морфологические признаки» описаны в рамках морфологической базы данных. Таблицы «Семантические значения» и «Типы семантических значений» описаны в рамках лексико-семантического словаря.

**Таблица «Предложные фреймы»**

Содержит описания предложных фреймов.

*Поля таблицы «Предложные фреймы»:*

* ID\_предложного\_фрейма – идентификатор предложного фрейма (Тип данных: целое число);
* ID\_лексема\_предлога – идентификатор лексемы предлога (Тип данных: целое число);
* ID\_сем\_значения\_до\_предлога – идентификатор дополнительного значения, которое может принимать слово расположенное перед предлогом (Тип данных: целое число);
* ID\_сем\_значения\_после\_предлога – идентификатор дополнительного значения, которое может принимать слово расположенное после предлога (Тип данных: целое число);
* ID\_падеж\_слова – идентификатор падежа, в котором должно находится слово, расположенное после предлога (Тип данных: целое число);
* ID\_сем\_значение\_фрейма – идентификатор семантического значения фрейма (Тип данных: целое число);
* Комментарий – необязательное поле, в котором может содержаться пример использования данного фрейма.

*Компонент разрешения имен*

В следующих разделах будет описана проблема неоднозначности именования параметров в онтологиях, обнаруженная при разработке алгоритма, и данный компонент содержит информацию необходимую для преодоления указанной проблемы при построении запроса. Схема этого компонента представлена на рисунке 5.

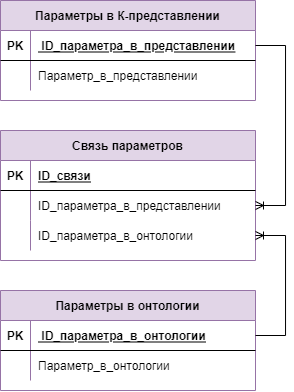


Рисунок 5 – Схема компонента разрешения имен параметров

*Описание таблиц компонента разрешения имен*

**Таблица «Параметры в К-представлении»**

Содержит список параметров, используемых в К-представлении.

*Поля таблицы «Параметры в К-представлении»:*

* ID\_параметра\_в\_представлении – идентификатор параметра, используемого в К-представлении (Тип данных: целое число);
* Параметр\_в\_представлении – параметр, используемый в К-представлении (Тип данных: строка переменной длины).

**Таблица «Параметры в онтологии»**

Содержит список параметров, используемых в онтологии.

*Поля таблицы «Параметры в онтологии»:*

* ID\_параметра\_в\_онтологии – идентификатор параметра, используемого в онтологии (Тип данных: целое число);
* Параметр\_в\_онтологии – параметр, используемый в онтологии (Тип данных: строка переменной длины).

**Таблица «Связи параметров»**

Содержит информации о связи между параметрами К-представления и онтологии.

*Поля таблицы «Связи параметров»:*

* ID\_связи – идентификатор конкретной связи параметров (Тип данных: целое число);
* ID\_параметра\_в\_представлении – идентификатор параметра, используемого в К-представлении (Тип данных: целое число);
* ID\_параметра\_в\_онтологии – идентификатор параметра, используемого в онтологии (Тип данных: целое число).

**Общая схема лингвистической базы данных**

Общая схема всей ЛБД представлена на рисунке 6.

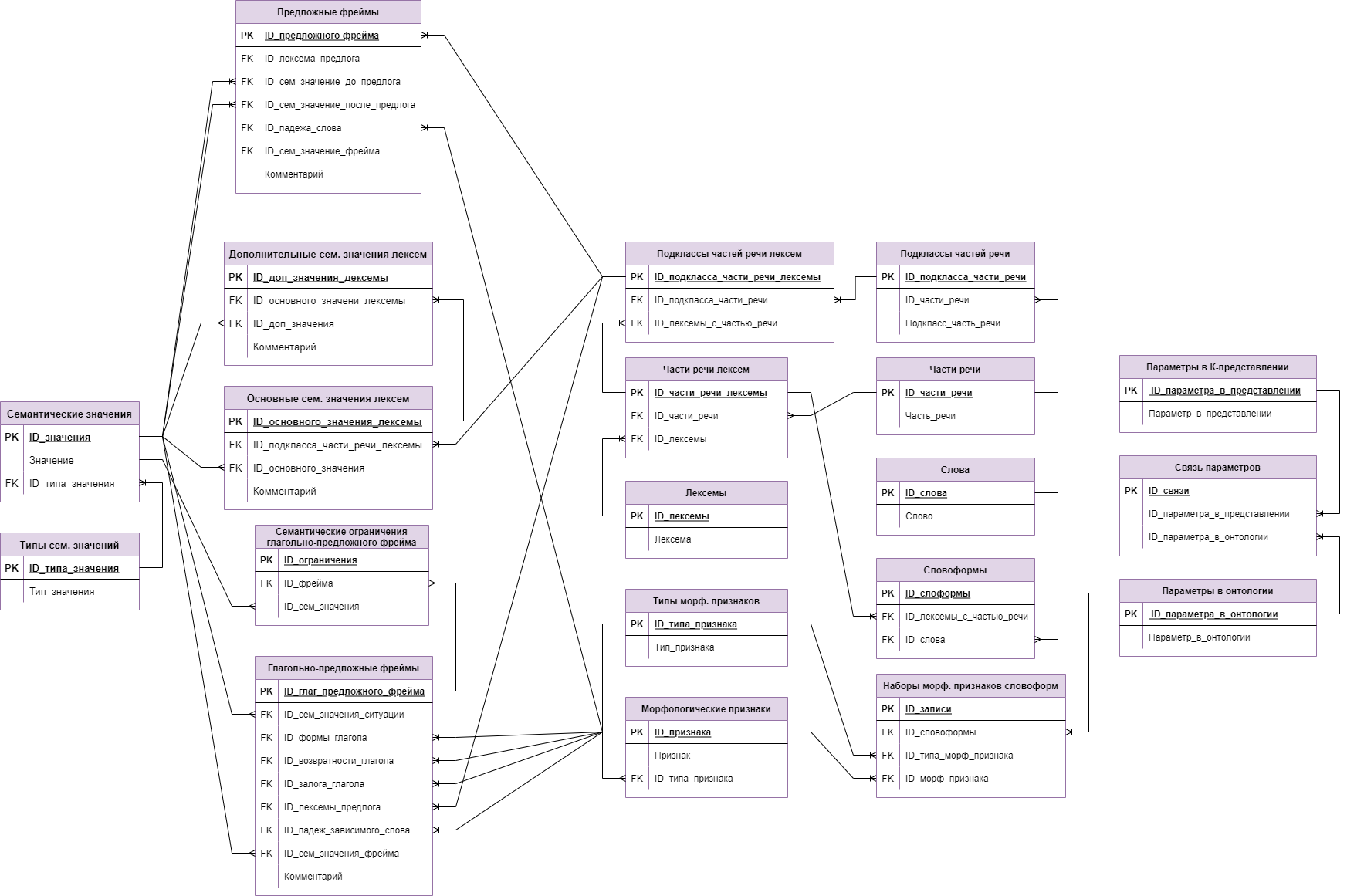


Рисунок 6 – Общая схема лингвистической базы данных

**Список литературы**

1. Linked Data // W3C URL: https://www.w3.org/DesignIssues/LinkedData.html (дата обращения: 10.02.2022).
2. RDF 1.1 Primer // W3C URL: https://www.w3.org/TR/rdf11-primer/ (дата обращения: 10.02.2022).
3. RDF Schema 1.1 // W3C URL: https://www.w3.org/TR/2014/REC-rdf-schema-20140225/ (дата обращения: 10.02.2022).
4. OWL 2 Web Ontology Language Primer (Second Edition) // W3C URL: https://www.w3.org/TR/2012/REC-owl2-primer-20121211/ (дата обращения: 11.02.2022).
5. Stadler C., Lehmann J., Höffner K., Auer S. LinkedGeoData: A Core for a Web of Spatial Open Data. Semantic Web 3(4):333-354. IOS Press, 2012. – 23 p. doi: 10.3233/SW-2011-0052 URL: https://www.researchgate.net/publication/240615213\_LinkedGeoData\_A\_Core\_for\_a\_Web\_of\_Spatial\_Open\_Data (дата обращения: 05.03.2022).
6. Tirad M. Almalahmeh, Sameem Abdul Kareem, Mansoor A. Abdulgabber Semantic recommender system with natural language interface: malaysian tourism industry // International Journal of Scientific & Engineering Research. - 2013. - Volume 4, Issue 9. - С. 1059-1065.
7. Bauer F., Kaltenböck M. Linked Open Data: The Essentials. A Quick Start Guide for Decision Makers. – 62 p. URL: https://www.reeep.org/LOD-the-Essentials.pdf (дата обращения: 05.03.2022).
8. BBK Linked Open Data // Официальный сайт Российской государственной библиотеки. URL: https://lod.rsl.ru/ (дата обращения: 15.03.2020).
9. SPARQL 1.1 Query Language // W3C URL: https://www.w3.org/TR/sparql11-query/ (дата обращения: 10.02.2020).
10. Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N. (2013).  Abstract Meaning Representation for Sembanking. *In: Proceedings of the 7th ACL Linguistic Annotation Workshop and Interoperability with Discourse*, Sofia, Bulgaria, August 8-9, 2013 (2013)(www.aclweb.org/anthology/W13-2322)
11. Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N. (2019).   *Abstract Meaning Representation (AMR) 1.2.6 Specification*; github.com/amrisi/amr-guidelines/blob/master/amr.md.
12. Montague Semantics // Stanford Encyclopedia of Philosophy URL: https://plato.stanford.edu/entries/montague-semantics/ (дата обращения: 16.03.2022).
13. Фомичев В.А. Формализация проектирования лингвистических процессоров. - М.: МАКС Пресс, 2005. - 367 с.
14. B. Nethravathi, G. Amitha, Anusha Saruka, T. P. Bharath, Setu Suyagya. Structuring Natural Language to Query Language: A Review // Engineering, Technology and Applied Science Research. - 2020. - №10(6). - С. 6521-6525.
15. Lee M. Christensen, Henk Harkema, Peter J. Haug, Jeannie Y. Irwin, Wendy W. Chapman (2009). ONYX: A System for the Semantic Analysis of Clinical Text
16. Chuan Wang, Nianwen Xue, Sameer Pradhan (2015). A Transition-based Algorithm for AMR Parsing.
17. Sylvain Pogodalla (2004). Computing Semantic Representation: Towards ACG Abstract Terms as Derivation Trees
18. Sébastien Ferré. SQUALL: a Controlled Natural Language for Querying and Updating RDF Graphs (2012).
19. Sébastien Ferré. SQUALL: A Controlled Natural Language as Expressive as SPARQL 1.1. (2013).
20. Neli Zlatareva, Devansh Amin. Natural Language to SPARQL Query Builder for Semantic Web Applications // Journal of Machine Intelligence and Data Science. - 2021. - Volume 2. - С. 44-53.