**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«Рязанский государственный радиотехнический университет»**

**(ФГБОУ ВО «РГРТУ», РГРТУ)**

«Утверждаю»

Заведующий кафедрой

ВПМ Пылькин А.Н.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  *(подпись)*

«\_\_\_» \_\_\_\_\_\_\_\_\_\_ 2017 г.

ОТЧЕТ

О ПРЕДДИПЛОМНОЙ ПРАКТИКЕ СТУДЕНТА

**4 курса, 344 группы**

*(Яровова О.С.)*

Рязань 2017

Направление подготовки: 09.03.03 «Прикладная информатика»

*(шифр и наименование)*

Тема ВКР: «Разработка информационной системы для представления цепочек преобразования типов формообразования естественного языка с помощью графов»

|  |  |  |
| --- | --- | --- |
| № | Результаты преддипломной практики студента | Результаты |
| 1 | Рассмотрение сферы исследования и обоснование её актуальности | Составлен отчет об актуальности представления цепочек преобразования типов формообразования естественного языка с помощью графов |
| 2 | Обоснование выбора темы | Составлен отчет о выборе темы |
| 3 | Изучение и анализ уровней автоматической обработки текста, а так же методов генерации и определения | Проведён анализ уровней АОТ и методов генерации и определения |
| 4 | Отчет о проделанной работе | Составлен отчет по результатам работы |

Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(Яровова О.С.)

*(Ф.И.О. студента, подпись)*

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2017 г.

Научный руководитель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(Пруцков А.В.)

*(Ф.И.О., подпись)*

*Студент четвертого курса Яровова О.С.*

*Направление 09.03.03. «Прикладная информатика»*

**Отчет об актуальности сферы исследования**

Под автоматической обработкой текста принято понимать процесс, направленный на преобразование текста на естественном языке при помощи ЭВМ. Эти преобразования используются для машинного перевода, анализа речи, ее синтеза, проверки правописания текстов, поиска и индексации информации в глобальных сетях и решения множества других задач.

Во все времена языковая грамотность является неотъемлемой частью любого образованного человека. Каждый из существующих языков обладает своими уникальными чертами и правилами, которые подчас вызывают сомнения в том, какими грамматическими характеристиками обладает слово, какую форму того или иного слова нужно употребить в определенном контексте и как получить необходимую форму. Для решения описанных задач используются информационные системы, которые содержат необходимую информацию.

Русская грамматика и словообразование являются одними из самых сложных, не зависимо от того является ли этот язык родным для изучающего или нет. Именно поэтому в настоящей работе рассматривается формообразование глаголов русского. Выбор данной части речи обусловлен многообразием ее форм и сложностей, связанных с ними, ведь правила русского формообразования глаголов сложны, разнообразны и богаты на исключения, именно поэтому для более эффективного изучения необходимо представление этапов формообразования необходимо осуществить наиболее наглядно.

Реализация этой задачи возможна при помощи представления преобразований в виде графа преобразований. Для его реализации необходимо осуществить классификацию глаголов по типам формообразования и описать получение цепочек словоформ последовательностью преобразований. После этого необходимо оптимизировать полученные цепочки, путем построения древовидной структуры, что позволит сократить скорость определения грамматического значения и основы. Полученные структуры дадут возможность программного получения из любой введенной формы глагола его грамматических характеристик и инфинитива, от которого образована рассматриваемая форма, также возможны и обратные преобразования - получение формы с необходимыми грамматическими характеристиками по инфинитиву.

**Целью выпускной квалификационной работы** является создание информационной системы по морфологии русского языка, позволяющей оптимизировать сложные и громоздкие последовательности преобразований, путем представления их в виде графа, а также позволяющей с помощью полученного графа получить необходимые грамматические характеристики определенной словоформы.

Справочная система по морфологии русского языка, на основе рассмотренного подхода, позволит:

* получить грамматические характеристики введенной глагольной формы;
* получить инфинитив глагола по одной из его форм;
* образовывать форму глагола в нужном виде, времени, лице, числе;
* образовывать особые формы глагола (причастия, деепричастия) в нужном виде, времени, лице, числе, падеже.

**Актуальность** темы выпускной квалификационной работы связана с распространенностью исследуемой проблемы и заключается в необходимости применения современных информационных технологий для оптимизации работы.Так как ручное формообразование является процессом, требующим затраты значительного количества временных ресурсов, а также глубоких познаний в области лингвистики, то создание информационной системы, позволяющей автоматизировать данный процесс, поможет избежать ошибок и потери времени в процессе определения грамматических характеристик рассматриваемой формы, а также образования словоформы с требуемыми грамматическими характеристиками от заданного инфинитива.

**Для достижения поставленных целей** в выпускной квалификационной работе реализованы следующие задачи:

* осуществлена классификация глаголов русского языка по типам формообразования;
* составлены полные парадигмы формообразования для каждого образованного типа;
* оптимизирован поиск необходимой парадигмы путем их объединения в граф преобразований;
* осуществлена разработка ПО для поиска грамматических характеристик по составленному графу.

**Практическая ценность.** Данная система может стать полезной для многих, чья жизнь в той или иной мере связана с русским языком. Школьникам она будет полезна для решения вопросов, связанных с трудностями, возникающими при выполнении домашних заданий: поиске инфинитива, определению грамматических характеристик слова. Иностранные студенты здесь смогут найти необходимую информацию, которая поможет им определить правильность употребления той или иной формы в построенной ими фразе на русском языке. Да и каждый человек сможет просто проверить себя, найдя интересующую его информацию.

Также возможна реализация метода и для других естественных языков. Справочная система может стать частью более крупных информационных ресурсов для автоматизации обучения и проверки знаний.

*Студент четвертого курса Яровова О.С.*

*Направление 09.03.03. «Прикладная информатика»*

**Обоснование выбора темы**

Основой грамотной письменной и устной речи каждого человека служит употребление верных словоформ в том или ином контексте. Каждый язык обладает своими правилами и особенностями формообразования, их изучение является процессом сложным и требует затраты большого количества времени. С другой стороны, для решения такой задачи можно прибегнуть к помощи словаря, но это также довольно трудоемкий процесс при необходимости поиска большого количества информации, осуществляемом вручную. Поэтому и возникла необходимость создания информационной системы, которая бы позволила получить необходимую словоформу, не прибегая к изучению правил формообразования и использованию бумажных словарей.

Информационная система по морфологии русского языка - это совокупность словаря цепочек преобразований, с относящимися к каждой из них инфинитивами, и программных средств и алгоритмов, позволяющих получить необходимую информацию о словоформе.

В настоящее время в сети Интернет существуют словари по морфологии русского языка, такие как «Морфология.ру», «udarenieru.ru». Однако они имеют недостатки. Однако, эти словари содержат информацию, разнящуюся с информацией из толково-грамматического словаря, например, Интернет словари говорят о том, что такие глаголы как *«портить, тратить»* имеют формы страдательного причастия настоящего времени, однако, в толково-грамматическом словаре содержится информация об обратном. Возможно, это связано с тем, что рассматриваемые Интернет словари составлены на основе грамматического словаря, который является изданным в 1977 г., тогда как, по информации из которого составлена разрабатываемая система, является более современным и издан в 1989 г. Так же Интернет словарями не представляется возможности воспользоваться без наличия сети Интернет, однако это вряд ли можно назвать недостатком, так как в наше время он окружает нас по всюду, и возможность им воспользоваться есть практически у каждого.

Для разработки информационной системы была выбрана среда разработки *Microsoft Visual Studio 2015.*

Среда разработки *Microsoft Visual Studio 2015* является популярной интегрированной средой разработки *(integrated development environment — IDE),* которая доступна в настоящее время программистам. Эта среда является результатом долгой истории развития языков программирования и интерфейсов и вобрала в себя достижения многих сред разработки программного обеспечения.

*Microsoft Visual Studio 2015* включает в себя редактор исходного кода с возможностью простейшего рефакторинга кода. Встроенный отладчик может работать как отладчик уровня исходного кода, так и отладчик машинного уровня. Остальные встраиваемые инструменты включают в себя редактор форм для упрощения создания графического интерфейса приложения, веб-редактор, дизайнер классов и дизайнер схемы базы данных. *Visual Studio* позволяет создавать и подключать сторонние дополнения (плагины) для расширения функциональности практически на каждом уровне, включая добавление поддержки систем контроля версий исходного кода, добавление новых наборов инструментов (например, для редактирования и визуального проектирования кода на предметно-ориентированных языках программирования) или инструментов для прочих аспектов процесса разработки программного обеспечения.

*Microsoft Visual Studio 2015* представляет собой высокоэффективную среду разработки, которая позволяет решать различные прикладные задачи, соединять несколько технологий, дающих разработчикам возможность создания таких приложений как *Windows*, *Web-служб*, *Web-приложений*, сайтов и т.д. *Microsoft Visual Studio 2015* позволяет обеспечить высокое качество кода на протяжении всего цикла разработки.

В качестве языка программирования был выбран язык *С#,* который является одним из самых современных языков программирования и встроен в среде *Microsoft Visual Studio 2015. С#* наследует большинство языковых конструкций языка *С++,* однако имеет и языковые конструкции, предназначенные для создания прикладного программного обеспечения, которые не имеют прямых аналогов в других языках*,* например, *деревья выражений, позволяющие представлять код в виде древовидной структуры данных, LINQ- (Language-Integrated Query)-* язык запросов к источнику данных, в качестве которого может выступать объект, реализующий интерфейс *IEnumerable* (например, стандартные коллекции, массивы), набор данных *DataSet,* документ *XML*. Но вне зависимости от типа источника *LINQ* позволяет применить ко всем один и тот же подход для выборки данных.

Так же для реализации поставленной задачи с помощью *Microsoft Visual Studio 2015* не требует материальных затрат, так как в рамках поставленной задачи достаточно бесплатной версии *Community*.

Таким образом, можно сделать вывод, что использование среды *Microsoft Visual Studio 2015* и языка *С#*, является оптимальным решением для выполнения поставленных задач.

*Студент четвертого курса Яровова О.С.*

*Направление 09.03.03. «Прикладная информатика»*

**Анализ уровней АОТ и методов генерации и определения**

Автоматическая обработка текста представляет собой преобразование текстов на естественном языке при помощи ЭВМ. Данная технология используется при анализе и синтезе речи, машинном переводе, проверке правописания текстов, а также для решения множества других задач.

Выделяют три уровня обработки текста:

* морфологический;
* синтаксический;
* семантический.

Морфологический анализ включает в себя выделение грамматической основы слова, определение принадлежности слова к той или иной части речи, приведение слова к словарной форме (инфинитиву), определение принадлежности словоформы слову путем анализа структуры словоформы.

Синтаксический анализ направлен на выявление связей между словами в предложении, построением синтаксической структуры предложения.

Семантический анализ характеризуется выявлением смысловых связей между словами и синтаксическими группами.

Каждый из уровней анализа является самостоятельной задачей, которая не имеет собственного практического применения, но является активно используемой для решения более общих задач. Многие исследовательские системы предназначены для решения именно таких задач. Эти системы применяются для апробации методов и проведения вычислительных экспериментов, либо в качестве составных частей (или библиотек) для систем, решающих ту или иную прикладную задачу. Можно сказать, что уровни автоматической обработки текста являются подсистемами общей системы естественного языка, они взаимосвязаны, но в тоже время достаточно автономные.

К первому этапу обработки текста относится морфологический анализ слов. Входным параметром является текстовое представление исходного слова. Целью и основным результатом морфологического анализа являются установление связи анализируемой словоформы с ее представлением в словаре, а также определение всех морфологических характеристик слова.

После того как проведен анализ каждого слова (морфологический анализ), начинается анализ предложений (семантический анализ), который позволяет определить взаимосвязь между отдельными словами и частями предложений. Результатом такого анализа является граф, узлы которого представляют собой слова предложения, при этом, если два слова связаны каким-либо образом, то соответствующие им вершины графа связаны дугой с определенной окраской. Возможные окраски дуг зависят от языка, на котором написано предложение, от выбранного способа представления синтаксической структуры предложения. Например, при синтаксическом анализе предложений русского языка в качестве окрасок дуг можно использовать вопросы, задаваемые от одного слова к другому. В вершинах графа слова пишутся не в том виде, в котором они встречаются в предложении, а в своей основной словоформе. Некоторым словам (например, предлогам) вообще не соответствует ни одна из вершин графа, но эти слова влияют на вопросы, задаваемые от одного слова к другому.

Семантический анализ в свою очередь основывается на результатах синтаксического анализа, в результате которого получается структура, отражающая синтаксическую структуру каждого предложения, а не набор слов, разделенных на предложения. Решение задач семантического анализа основываются на результатах анализа отдельных слов, и вместо синтаксической структуры предложения анализу подвергаются наборы стоящих рядом слов. Так как методы семантического анализа работают со смыслом слов, то необходим универсальный набор семантических единиц при помощи которого можно было бы выразить смысл любого высказывания. На данном этапе основой является словарь синонимов, антонимов, омонимов, паронимов и т.д.

Все уровни обработки являются взаимосвязанными, и чтобы преобразовать текстовой запрос пользователя в его смысловое представление и, обратно, преобразовать смысл ответа в его текстовое представление необходимо выполнить автоматическую обработку текста на морфологическом, синтаксическом и семантическом уровнях.

В основе проектируемой информационной системы лежит морфологический уровень автоматической обработки текста. На этом уровне осуществляется морфологический анализ и синтез.

При морфологическом анализе для каждого слова осуществляется задача определения - нахождения по данной словоформе ее нормальной формы (инфинитива) и грамматического значения.

Морфологический синтез заключается в решении для каждого слова задачи генерации. Генерация формы слова - это процесс получения необходимой словоформы по известному инфинитиву и грамматическому значению.

Для морфологической обработки текста используются словари, в которых содержатся классифицированные слова. Чтобы классифицировать слова, необходимо построить полную парадигму, то есть совокупность всех словоформ слова.

В проводимой классификации рассматривается получение словоформ слова с помощью двух основных методов преобразования:

* добавление подстроки к строке слева или справа;
* замена первой слева подстроки на другую подстроку.

Каждое основное преобразование имеет обратное к нему преобразование. Все остальные преобразования возможно представить в виде конечной последовательности перечисленных основных и/или обратных типов преобразований. Данные понятия используются в универсальной модели формообразования.

В качестве естественного языка в рассматриваемой системы был выбран русский язык. Данный выбор обусловлен тем, что этот язык является богатым на различные словоформы, например, в отличие от многих других в русском языке возможно изменение слов по падежам.

Для более детального рассмотрения и классификации был выбран глагол- часть речи русского языка, обозначающая действия или состояния разного характера.

Глагол обладает различными грамматическими категориями и соответствующими словоформами. Как и другие части речи, он имеет категории числа, рода, лица, однако есть и уникальные, присущие только глаголу- категории вида, времени, залога, наклонения. Таким образом, глагол является самой сложной, с точки зрения, формообразования частью речи русского языка, так как он обладает уникальными признаками, присущими лишь глаголу, поэтому была проведена классификация именно этой части речи.

Все глагольные формы делятся на спрягаемые и неспрягаемые.

К спрягаемым относятся глаголы настоящего, будущего и прошедшего времени изъявительного наклонения, в первом, втором и третьем лице, единственном и множественном числе. А также глаголы повелительного, в единственном и множественном числе, и условного наклонений, в первом, втором и третьем лице, единственном и множественном числе.

К неспрягаемым формам относятся отглагольные формы глагола, деепричастия совершенного и несовершенного вида, а также причастия настоящего и прошедшего времени, действительного и страдательного залога, женского, мужского и среднего родов, а также именительного, родительного, винительного, дательного, творительного и предложного падежей.

Для проведения классификации был выбран толково-грамматический словарь глаголов русского языка.

Процесс классификации включает в себя следующие этапы:

* составить полную парадигму словообразования слова;
* сравнить составленную цепочку с уже существующими;
* при полном совпадении цепочек преобразований отнести рассматриваемое слово к уже имеющемуся типу;
* при отсутствии полного совпадения, образовать новый тип;
* типом называется группа, к которой относятся два и более слова, в противном случае, член типов относят к группе исключений - слов, которые изменяются индивидуальным образом.

В результате классификации было рассмотрено 2500 инфинитивов, которые разделились на 124 типа, к глаголам, изменяемым индивидуальным образом отнесено 134.

Такое многообразие типов связано с тем, что русский является богатым на особенности орфографии, а также с тем, что многие глаголы, имеющие одинаковые парадигмы в глагольных формах, отличаются наличием различных отглагольных форм – причастий, например, глагол может иметь только действительные причастия настоящего и прошедшего времени, или действительные причастия настоящего и прошедшего времени, и страдательные причастия настоящего времени, или действительные причастия настоящего и прошедшего времени, и страдательные причастия прошедшего времени, или действительные и страдательные причастия настоящего и прошедшего времен.

Для решения поставленных задач используются алгоритмы генерации и определения.

Алгоритм определения заключается в определении грамматического значения и основы (инфинитива) рассматриваемой словоформы. Суть алгоритма заключается в последовательном переборе цепочек преобразований типов. Если последовательность преобразований применима к рассматриваемой словоформе и тип цепочки совпадает с типом полученной основы в ходе совершенных преобразований, то такие цепочки называются целевыми, в противном случае – ложными. Так же существуют тупиковые цепочки – цепочки, выполнение преобразований которых невозможно к рассматриваемой словоформе, что приводит к ее прерыванию. При нахождении первой целевой цепочки процесс поиска не заканчивается, это связано с существованием омонимии в естественных языках, таким образом не одна, а несколько цепочек могут быть применимы к рассматриваемой словоформе. Так, например, к словоформе *золотящий* может быть применена последовательность преобразований *действительного причастия настоящего времени единственного числа мужского рода как в именительном падеже, так и в винительном падеже для неодушевленной формы*, обе эти цепочки будут являться целевыми.

В свою очередь алгоритм генерации заключается в поиске необходимой последовательности преобразований по имеющейся основе (инфинитиву) и грамматическому значению, и применении ее для получения необходимой словоформы.

Цепочки преобразований можно представить в виде графа. Вершинами графа являются словоформы, корневая вершина – основа, промежуточные – результаты преобразований, конечные вершины – словоформы. Дуги графа – необходимые преобразования, которые необходимо осуществить, чтобы получить из одной формы другую. Путь в графе представляет собой цепочку преобразований, при этом для задачи генерации используется путь от корневой вершины к конечной – прямая цепочка преобразований, а для задачи определения – обратная, заключающаяся в движении от конечной вершины к корневой, то есть нахождения такой конечной вершины, которая приедет к корневой.

Процесс движения по графу в задаче определения состоит из следующих действий:

* применить преобразование к словоформе;
* проверить применимость преобразования к словоформе: если преобразование применимо – перейти в следующее состояние, иначе считать путь тупиковым;
* по достижении конечной вершины проверить совпадение типов цепочки и полученной основы, считать путь целевым, иначе – ложным.

Процесс движения по графу в задаче генерации заключается в:

* определении типа основы;
* определении цепочки преобразований, соответствующей заданному грамматическому значению и типу основы;
* применении найденной цепочки к основе.