МИНОБРНАУКИ РФ
Федеральное государственное бюджетное образовательное
учреждение высшего профессионального образования
«Ижевский Государственный Технический Университет имени М.Т.Калашникова»
Кафедра «АСОИУ»

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

к дипломному проекту по специальности 230100
«Автоматизированные системы обработки информации и управления»
на тему:

«Разработка метода и программной системы морфемного анализа слов русского языка»

Проектант
студент гр. 8-78-5 Е.А. Черных

Руководитель дипломного проекта
к.т.н., доцент каф. АСОИУ М.Н.Мокроусов

Нормоконтроль
ст. преп. каф. АСОИУ Н.В. Соболева

Заведующий кафедрой АСОИУ
д.т.н., профессор В.Н. Кучуганов

Декан факультета ИВТ
д.т.н., д.э.н., д.г.-м.н В.Е. Лялин

Ижевск 2013

РЕФЕРАТ

Дипломный проект на тему «Разработка метода и программной системы морфемного анализа слов русского языка» содержит 88 страниц текста, 42 иллюстрации, 34 таблиц, 3 приложения. В ходе дипломного проектирования было использовано 23 литературных источников.

Целью дипломной работы является повышение степени применимости средств анализа русскоязычного текста в задачах автоматизированного обучения русскому языку Основные результаты:

1 Проведен аналитический обзор методов и систем морфемного анализа русскоязычных текстов;

2 Разработана математическая модель морфемного анализа слов русского языка;

3 Разработана функциональная модель системы и ее структуры;

4 Разработаны подсистемы морфологического анализа слов русского языка на основе библиотеки MCR.dll и программы mystem;

5 Разработана база данных, необходимая для морфемного анализа слова;

6 Разработан алгоритм сокращения вариантов морфемного разбора;

7 Разработана подсистема экспертного редактирования базы данных системы;

8 Разработана подсистема визуализации результатов морфемного разбора;

9 Спроектирована и разработана программная система морфемного анализа слов русского языка.

В результате выполнения дипломной работы был разработан и создан программный продукт, автоматизирующий морфемный анализ слов русского языка. Исследованы принципы морфемного строения слов, разобраны проблемы автоматизации анализа текстов на естественном языке в рамках изученной темы. Намечены новые цели и задачи, для повышения степени эффективности разработанной программной системы.

СОДЕРЖАНИЕ

СПИСОК ИСПОЛЬЗУЕМЫХ СОКРАЩЕНИЙ 4
ВВЕДЕНИЕ 5
1 Аналитический обзор методов и систем морфемного анализа русскоязычных текстов 7
2 Морфемы русского языка и правила морфемного анализа 11
2.1 Корень, основа слова и окончание 11
2.2 Приставки 12
2.3 Суффиксы 14
2.4 Примеры морфемного разбора слов разных частей речи 23
3 Математическая модель морфемного анализа слов русского языка 25
4 Функциональная модель системы и её структура 27
5 Проектирование программной системы 31
5.1 Разработка подсистемы морфологического анализа слов русского языка на основе библиотеки MCR.dll и программы mystem 31
5.2 Разработка алгоритма сокращения вариантов морфемного разбора 33
5.3 Разработка подсистемы визуализации результатов морфемного разбора 36
5.4 Разработка подсистемы экспертного редактирования базы данных системы 38
6 Проектные решения по системе 43
6.1 Выбор и обоснование программных и технических средств 43
6.2 Логическая модель базы данных 45
6.3 Физическая модель базы данных 51
6.4 Создание запросов 53
7 Тестирование системы 63
8 Проблемы и задачи, выявленные в ходе разработки системы 67
ЗАКЛЮЧЕНИЕ 72
СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ 73
Приложение А (обязательное) Руководство пользователя 75
Приложение Б (обязательное) Текст программы 82
Приложение В (справочное) Таблицы переменных и постоянных грамматических характеристик 83

СПИСОК ИСПОЛЬЗУЕМЫХ СОКРАЩЕНИЙ

БД	− База данных
Сущ.	− Существительное
Прилаг.	− Прилагательное
Несов. в.	− Несовершенный вид
Им.п	− Именительный падеж
Род.п	− Родительный падеж
Ед.ч	− Единственное число
Мн.ч	− Множественное число
М.р	− Мужской род
Прош.вр	− Прошедшее время
Наст.вр	− Настоящее время
3 л	− 3 лицо

ВВЕДЕНИЕ

Морфемный анализ, или иначе разбор слова по составу – это действие, направленное на выделение в слове его минимальных значимых частей (морфем).

Целью системы морфемного анализа слов русского языка является автоматизация разбора слов по составу. Её создание является актуальной задачей для обучающих систем русского языка, так как визуализация результатов морфемного разбора, объяснение результатов морфемного анализа, позволит упростить изучение данного раздела языкознания.

В настоящее время в информационных системах на первый план выходит задача построения различного рода систем, которые выполняли бы роль конвертера человеческой речи в компьютерный формализованный язык. В связи с этим, одной из важных задач является разработка системы автоматического анализа текста.

Помимо этого накопленная информация о морфемном составе слов, о принципах строения слов может быть использовано в системах искусственного интеллекта. Данная информация позволила бы проанализировать новые слова, поступающие в систему. Строить аналогии их происхождения, объяснять правописание. В идеале изучение состава слова, исследование принципов соединения морфем в слово обязано служить овладению изучаемым языком, накопленная информация – приводить к анализу языка, на основе новых слов, появляющихся в нем.

На данном этапе разработки система рассчитана на изучение морфемного анализа в школах и ВУЗах, она может быть внедрена в автоматизированные системы обучения русскому языку, использоваться для генерации заданий по морфемике на основе произвольного текста, а также для объяснения результатов морфемного анализа.

Задачи, поставленные для реализации системы следующие:

10 Провести аналитический обзор методов и систем морфемного анализа русскоязычных текстов;

11 Разработать математическую модель морфемного анализа слов русского языка;

12 Разработать функциональные модели системы и ее структуры;

13 Разработать подсистемы морфологического анализа слов русского языка на основе библиотеки MCR.dll и программы mystem;

14 Разработать базу данных, необходимую для морфемного анализа слова, хранящую:

– морфемы и их признаки;

– морфемные словари и результаты разбора слова;

15 Разработать алгоритм сокращения вариантов морфемного разбора;

16 Провести проектирование программной системы;

17 Разработать подсистему экспертного редактирования базы данных системы;

18 Разработать подсистему визуализации результатов морфемного разбора;

19 Провести программную реализацию системы.

1 Аналитический обзор методов и систем морфемного анализа русскоязычных текстов

Введем основные понятия для рассмотрения данной темы.

Термин ***морфемика*** имеет два значения: 1. Морфемный строй языка, совокупность вычленяемых в словах морфем и их типы. 2. Раздел языкознания, изучающий типы и структуру морфем, их отношения друг к другу и к слову. Предметом изучения морфемики как раздела языкознания являются морфемы, их конкретные представители в слове — морфы, и их сочетания в слове [1].

Классификация морфем представлена на рисунке 1.1 [2].

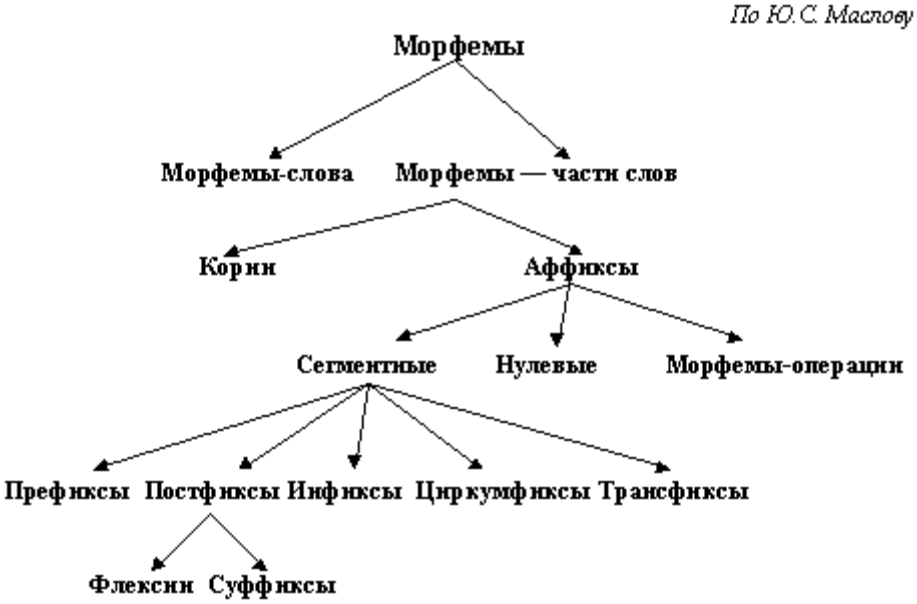


Рисунок 1.1 – Классификация морфем

Морфемный анализ (разбор слова по составу) – это действие, направленное на выделение в изучаемом слове его минимальных значимых частей (морфем).

При морфемном разборе слова сначала в слове выделяется окончание и формообразующий суффикс (если они есть), подчёркивается основа [3].После этого основа слова разбивается на морфемы. Существует два метода членения основы слова на морфы: формально-структурный и формально-смысловой.

Суть формально-структурного морфемного разбора состоит в том, что в основе в первую очередь выделяется корень как общая часть родственных слов. Затем то, что идёт до корня, должно быть осознано как приставка (приставки) в соответствии с представлениями о том, встречались ли подобные элементы в других словах. Аналогично с суффиксами. Иначе

говоря, главным при разборе становится эффект узнаваемости морфем, внешнее сходство каких-то частей разных слов.

Главной установкой формально-смыслового метода и алгоритм морфемного разбора выходят из трудов Г.О.Винокура [3] и состоят в неразрывности морфемного членения и словообразовательного разбора. Алгоритм морфемного разбора основы состоит в построении словообразовательной цепочки «наоборот»: со слова как бы «снимаются» приставки и суффиксы, корень же выделяется в последнюю очередь.

Таким образом, порядок разбора слова по составу таков:

1) выделить окончание, формообразующий суффикс (если они есть в слове);

2) выделить основу слова – часть слова без окончаний и формообразующих суффиксов;

3) выделить в основе слова приставку и/или суффикс через построение словообразовательной цепочки;

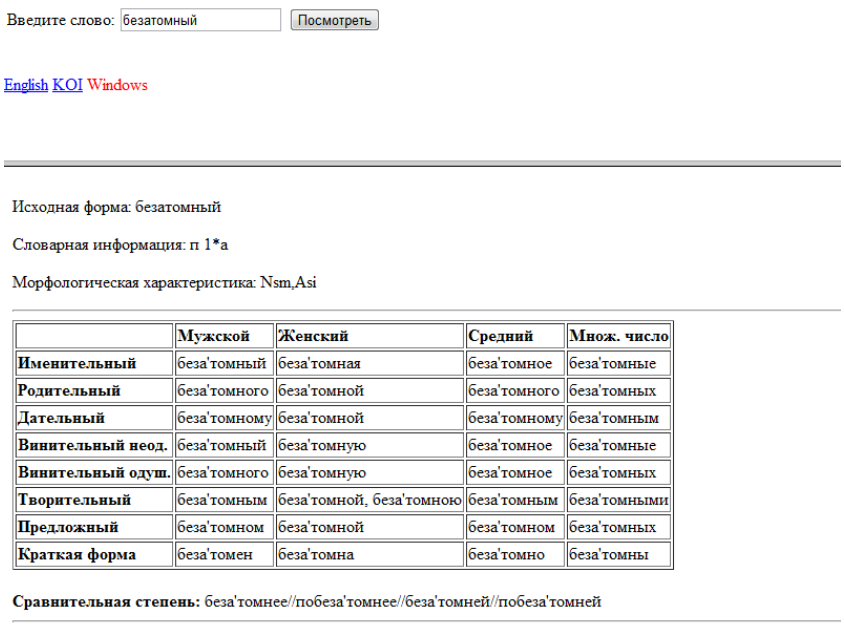
4) выделить в слове корень.

Существуют морфемно-орфографические словари, в которых приведены строения слов русского языка, например: «Школьный словообразовательный словарь» З.А. Потиха, «Морфемно-орфографический словарь» Тихонов А. Н и другие. Для некоторых созданы электронные версии словарей, доступные в Интернете. Например, рассмотрим разбор нескольких слов, приведенных в электронном словаре Тихонова А.Н.[4].

Разбор слова *безатомный* приведен, как ***Без/а́том/н/ый***, разбор слова *глазеть* – ***Глаз/е́/ть***, *яблоко* – ***я́блок/о***. Таким образом, в словах расставлены ударения и границы между морфами разделены слэшем (/). К какой конкретной морфеме относится каждая часть слова не указывается, это является главным недостатком данных словарей.

Что касается автоматизированных систем морфемного разбора слов русского языка, то подобных систем найдено не было. Их создание является актуальной задачей для обучающих систем русского языка. Если бы система визуализировала результаты морфемного разбора, объясняла результаты морфемного анализа и допущенные орфографические ошибки, то изучение данного раздела языкознания значительно бы упрощалось.

Существуют предпосылки для решения данной проблемы, так в свободном доступе можно найти базу данных корней русских слов (включая иностранные заимствования) [5]. Значительно продвинулись разработки систем морфологического анализа слов, их использование может помочь в выделении окончаний и основ слов. Пример, морфологический анализатор С.А.Старостина, основанный на словаре Зализняка А.А. [6]. На рисунке 1.2 и 1.3 представлен морфологический разбор слов *безатомный* и *яблоко*.


Введите слово: безатомный Посмотреть

English KOI Windows

Исходная форма: безатомный

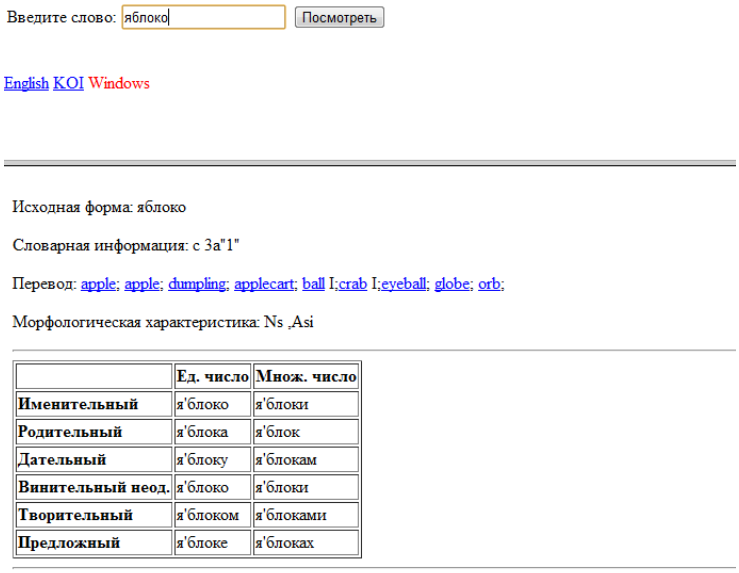
Словарная информация: п 1*а

Морфологическая характеристика: Nsm,Asi

	Мужской	Женский	Средний	Множ. число
Именительный	беза'томный	беза'томная	беза'томное	беза'томные
Родительный	беза'томного	беза'томной	беза'томного	беза'томных
Дательный	беза'томному	беза'томной	беза'томному	беза'томным
Винительный неод.	беза'томный	беза'томную	беза'томное	беза'томные
Винительный одуш.	беза'томного	беза'томную	беза'томное	беза'томных
Творительный	беза'томным	беза'томной, беза'томною	беза'томным	беза'томными
Предложный	беза'томном	беза'томной	беза'томном	беза'томных
Краткая форма	беза'томен	беза'томна	беза'томно	беза'томны

Сравнительная степень: беза'томнее//побеза'томнее//беза'томней//побеза'томней

Рисунок 1.2 – Морфологический разбор слова *Безатомный*


Введите слово: яблоко Посмотреть

English KOI Windows

Исходная форма: яблоко

Словарная информация: с 3а"1"

Перевод: apple; apple; dumpling; applecart; ball I;crab I;eyeball; globe; orb;

Морфологическая характеристика: Ns ,Asi

	Ед. число	Множ. число
Именительный	я'блоко	я'блоки
Родительный	я'блока	я'блок
Дательный	я'блоку	я'блокам
Винительный неод.	я'блоко	я'блоки
Творительный	я'блоком	я'блоками
Предложный	я'блоке	я'блоках

Рисунок 1.3 – Морфологический разбор слова *Яблоко*

На основе сравнения неизменяемой части слова возможно выделение основы и окончания слов, в приведенных примерах это: основа слова *безатомный* – *безатомн*, окончание – *ый*, в слове *яблоко* основа – *яблок*, окончание – *о*.

Минус систем, основанных на словаре А.А. Зализняка, заключается в ограниченности слов, которые могут быть ими проанализированы, количеством слов словаря. Данное ограничение может быть снято при использовании системы mystem от Яндекс [7]. Программа mystem производит морфологический анализ текста на русском языке. Для слов, отсутст-

вующих в словаре, порождаются гипотезы. Используя данную программу, так же можно производить выделение основы и окончания слова.

Кроме всего прочего, правила русского языка позволяют сгруппировать морфы в зависимости от части речи в которых они могут встречаться, так для глаголов свойственны такие окончание, как: *-у, -ю, -а, -и, -ешь, -ет, -ем, -ете, -ут, -ют, -ишь, -ит, -им, -ите, -ей, -ят, -ть, -ся* [8] и др. Из чего можно сделать вывод, что определив часть речи и распределив морфы по группам, значительно проще будет производить морфемный разбор слова.

Таким образом, разработка программной системы морфемного анализа слов русского языка трудная алгоритмическая задача, имеющая достаточное количество предпосылок для ее решения.

2 Морфемы русского языка и правила морфемного анализа

2.1 Корень, основа слова и окончание

Окончанием слова называется его изменяемая часть, которая служит для связи слов и выражает значения рода, числа, падежа, лица [8].

Окончания выражают значения:

1) рода, числа и падежа – у существительных и прилагательных;

2) лица и числа – у глаголов в настоящем времени;

3) рода и числа – у глаголов в прошедшем времени, у кратких прилагательных.

Окончания есть только у изменяемых слов. Нет окончаний у служебных слов, наречий, неизменяемых существительных и прилагательных. У изменяемых слов нет окончаний в тех их грамматических формах, в которых отсутствуют указанные грамматические значения (род, лицо, число, падеж), то есть у инфинитива и деепричастия.

Окончание может быть нулевым. Нельзя путать слова с нулевым окончанием и слова, в которых нет окончаний, — неизменяемые слова. Нулевое окончание не выражено звуком и на письме не обозначено буквой. Нулевые окончания обнаруживаются [3]:

1) у существительных в форме именительного падежа, единственного числа, мужского рода (2 склонения) и женского рода (3 склонения);

2) у части существительных в форме родительного падежа, множественного числа;

3) у кратких прилагательных и причастий в форме единственного числа, мужского рода;

4) у глаголов в форме прошедшего времени, единственного числа, мужского рода;

5) у притяжательных прилагательных с суффиксом *-ий-*;

6) у глаголов в повелительном наклонении, где нулевым окончанием выражается значение единственного числа.

Часть слова без окончания называется основой. Основа слова может быть равна корню. Кроме корня основа может включать приставки, суффиксы, соединительные гласные.

Обязательной частью основы любого слова является корень. Слова с несколькими корнями называются сложными. Части основы сложного слова могут быть соединены соединительными гласными.

Корень — это главная значимая часть основы, в нем заключено общее лексическое значение всех слов с этим же корнем (однокоренных, или родственных). Чтобы найти корень слова, надо подобрать к этому слову родственные слова.

2.2 Приставки

Приставка или префикс (от лат. prae — впереди, fixus — прикреплённый) – значимая часть основы [9]. Служит для образования новых слов.

С помощью приставки, в отличии от суффиксов, образование слов обычно идет в пределах одной и той же части речи [9]. Приставочным методом чаще всего образовываются наречия, прилагательные и глаголы. Приставкам несвойственна закрепленность за определенными частями речи в связи с большой степенью отвлеченности и универсальности выражаемых ими значений.

В русском языке насчитывается несколько десятков приставок (таблица 2.1) [10].

Таблица 2.1 – Приставки в русском языке

Приставка	Значение
А- / ан-	греческого происхождения (передаёт отрицание). Аналог русских приставок «не-», «без-».
Анти-	греческого происхождения (гр. «против»). Аналог русской приставки «противо-».
Безо- / без-/бес-	«за исключением», «кроме»
Во- / в-	внутрь, внутри
Возо- / воз- / вос- / взо- / вз-	Вверх
Взаимо-	Обоюдно
Вне-	Снаружи
Вы-	из, изнутри
До-	передает направление
Еже-	каждый. Аналог слова-приставки «каждо-»
За-	сзади, за спиной, за спину
Ино-	Другой
Любо-	слово-приставка, используется редко, в именах и другом словосложении: *любомудрие* - философия
Между- / меж-	слово-приставка: *международный*
Мимо-	слово-приставка: *мимоходом*
Миро-	слово-приставка, используется редко, в именах и другом словосложении: *Мирослав, миропорядок*
На-	наверх, наверху
Надо- / над-	сверху (не прикасаясь).

Продолжение таблицы 2.1

Приставка	Значение
Обо- / об- / о-	вокруг, кругом
По-	Вдоль
Подо- / под-	внизу, ниже низа предмета
Пере- / пре	через верх
При-	рядом, также передаёт приближение
Про-	Сквозь
Противо- / против-	слово-приставка. В русском языке в этом значении используются также аналогичные греческая приставка «анти-» (*антинародный*) и латинская приставка «контр-» (*контр-наступление*).
Со- / с-	быть вместе, делать совместно
Со- / с-	спускать(ся) вниз
Среди-	слово-приставка со значением «по-середине»: *срединземный*
Су-	быть рядом: *супруги, суглинок, супесь*
Тре-	избыточно, слишком много: *треволнения*
У-	Отдаление
Через- / чере- рес- / чрез-	Слишком
Эу-, эв-	греческого происхождения (гр. «истинно, правильно»), аналог слов-приставок «истинно-», «благо-»: евангелие («благовествование» - христианские книги Нового завета различных авторов), евхаристия («благодарение» - христианский обряд причащения), эуархонтоглиры («истинные пращуры»), *эубактерии* («настоящие бактерии»).

2.3 Суффиксы

Суффикс (от лат. Suffixus «прикреплённый») – значимая часть основы, разновидность аффиксов [11]. Служит для образования новых слов. Чаще всего стоит после корня.

В основном суффиксы свойственны определенным частям речи. Суффиксы существительных приведены в таблице 2.2, прилагательных – в таблице 2.3, глаголов – таблица 2.4, причастий – таблица 2.5, деепричастий – таблица 2.6, наречий – таблица 2.7 [11].

Таблица 2.2 – Суффиксы существительных

Суффиксы субъективной оценки	
Суффикс	Значение
-ашк-, -ишк-, -ышк-, -ушк-, -юшк-	*старикашка, голосишко, крылышко, травушка, волюшка*
-ек, -ик, -ок, -чик	*замочек, ключик, снежок, графинчик*
-ец-, -иц-	*морозец, книжица*
-ечк-, -ичк-, -очк-	*сестричка, утречко, звездочка*
-еньк-, -оньк-	*лисонька, зоренька*
-енк-, -инк-, -онк-	*снежинка, лошаденка, книжонка*
-ин-	*домина*
-ищ-	*голосище*
-к-	*головка, ночка*
-ушек, -ышек	*воробушек, колышек*
Словообразовательные	
-ак (-як)	существительные со значением: • лица, характеризующегося отношением к предмету, с которым связана его профессия (*моряк, рыбак*), или к названию государства, страны, местности, города, жителем которого данное лицо является (*земляк, пермяк, сибиряк*); • лица, характеризующегося признаком, названным мотивирующим именем прилагательным и определяющим его внешние качества, характер, родство, социальное положение (*бедняк, левак, пошляк, свояк, толстяк, чужак*); • лица, характеризующегося действием, названным мотивирующим глаголом (*вожак, чудак*)

Продолжение таблицы 2.2

Суффикс	Значение
-ан (-ян)	существительные со значением: • части тела, характеризующейся интенсивным внешним признаком (*пузан, лобан, губан*); • лица, склонного к тому, что названо исходным словом (*интриган, политикан, критикан, грубиян*); • названия животного (*орлан*); • явления, характеризующегося отношением к тому, что названо исходным словом (*буран*)
-анин (-янин)	существительные со значением лица по месту жительства (*горожанин, селянин*)
-ач	существительные со значением: • лица по преобладающему признаку (*силач, усач, трубач*); • предмета, который служит для выполнения действия (*тягач, пугач*).
-ев-, -ив-	от отглагольных существительных или от глаголов: *варево, месиво*
-евич, -евн-, -ович, -овн-, -ич, -иничн-, -ичн-	*Петрович, Петровна, Сергеевич, Сергеевна, Фомич, Ильинична, Никитична*
-ени[й-э] (-ни[й-э])	существительные со значением действия (*цветени[j-э], формировани[j-э], спасени[j-э]*)
-ет-(-от-)	существительные со значением: • отвлеченного признака (*быстрота, доброта, нищета*); • совокупности лиц (*беднота, пехота*); • действия с общим значением шума (*грохот, топот*).
-еств-(-ств-)	существительные со значением: • союза, объединения лиц, которые названы в исходном слове (*братство, начальство, юношество, землячество*); • со значением учреждения (*посольство, представительство*); • отвлеченного признака (*лукавство, богатство, изящество, супружество*); • лица, наделенного признаком, названным исходным словом (*божество, ничтожество, высочество*)
-есть (-ость)	существительные со значением отвлеченного признака или состояния (*свежесть от свежий, бледность от бледный, смелость от смелый, жалость*)

Продолжение таблицы 2.2

Суффикс	Значение
-ец	существительные со значением • лица по принадлежности к стране, территории, городу, где оно проживает или откуда происходит (*испанец, новгородец, горец*); • лица, характеризующегося каким-либо свойством (*мудрец, глупец, упрямец*); • предмета или явления, характеризующегося признаком или действием, названным словами, от которых они образованы (*резец, рубец, холодец*).
-изм	существительные, обозначающие состояния, качества, названия учений и общественных течений (*реализм, фанатизм, романтизм, героизм*).
-изн	существительные со значением: - отвлеченного признака (*белизна, желтизна, новизна*); - глагола со значением действия (*укоризна*)
-ик (-ник)	существительные, обозначающие: • лицо по свойству или признаку, которые определяют его отношение к предмету, занятию (*целинник, химик, очник*) • предмет, предназначенный для чего-либо (*чайник, приемник, бумажник*); • предмет, обозначающий книгу или сочинение (*задачник, справочник*); • пространство или территорию, покрытые чем-то или содержащие что-то (*ельник, малинник*); • вареник (*от вареный*), современник (*от современный*)
-ин	существительные со значением: • видов рыбы и мяса (*осетрина, баранина*); • одного предмета из ряда одинаковых (*горошина, соломина*); • отвлеченных признаков (*глубина, седина*); • результата или орудия действия (*перекладина, отметина, царапина*); • лица, которое является представителем какой-либо нации, гражданином какого-либо государства, жителем или уроженцем какой-либо страны (*грузин, татарин, мордвин*)
-ист	существительные со значением лица по принадлежности к учреждению, профессии, определенному общественному направлению (*связист, баянист, марксист*)

Продолжение таблицы 2.2

Суффикс	Значение
-иц- (-ниц-)	существительные со значением: • предмета - вместилища чего-нибудь, места (*сахарница, больница*); • названия самок животных (*волчица, львица*) • лиственница (*от лиственный*) • писательница (*от писатель*)
-их-	существительные - названия самок животных (*зайчиха, слониха, кроль-чиха*)
-к-	существительные, обозначающие: • лиц женского пола от соответствующих имен существительных муж-ского (*спортсменка, студентка*); • предмет (машину, приспособление, орудие, помещение), предназна-ченный для осуществления действия (*задвижка, терка, перегородка*); • предмет - результат действия (*записка, настойка*); • действие (*вспышка, попытка*).
-л-	существительные со значением: • лица, которое постоянно или обычно выполняет действие (*зубрила, громила, вышибала*); • предмета, предназначенного для осуществления действия (*мыло, под-дувало*); • предмета или явления, характеризующихся действием, названным ис-ходным словом (*быль, прибыль, поросль*).
-лк-	существительные со значением: • предмета, предназначенного для выполнения действия (*сеялка, грелка, копилка*); • помещения, предназначенного для осуществления действия (*парилка. раздевалка*); • названия лиц, выполняющих действия (*гадалка, сиделка*).
-льн-	• существительные со значением места совершения действия (*гладиль-ня, купальня, спальня*); • прилагательные со значением предназначенности для выполнения действия (*вязальный, родильный, сушильный*).
-льник-	существительные со значением предмета, предназначенного для вы-полнения действия (*холодильник, умывальник*).
-льщик- (-льщиц-)	существительные со значением лица по роду деятельности, занятию или действию (*бурильщик, точильщик, чистильщик*)

Продолжение таблицы 2.2

Суффикс	Значение
-тель (-итель)	существительные со значением: • лица, принадлежащего к той или иной профессии, занимающегося той или иной деятельностью (*учитель, искатель, воспитатель, спасатель*); • предмета (орудия, приспособления, машины), который производит действие (*двигатель, выключатель, глушитель*).
-ун	существительные со значением: • лица по действию, характерному для него (*бегун, крикун*); • животных по характерному для них признаку (*грызун, скакун*)
-чик (-щик, -чиц-)	существительные со значением: • лица по роду деятельности (*бетонщик, переводчик, жестянщик, буфетчик*); • предмета (машины, механизма, приспособления), который производит действия (*погрузчик, буксировщик*)

Таблица 2.3 – Суффиксы прилагательных

Суффикс	Значение
-ал- (-ел-)	прилагательные со значением такой, каким становятся под влиянием действия (*лежалый, загорелый, устарелый*)
-ан- (-ян-)	прилагательные со значением: • сделанный из того или иного материала или относящийся к чему-то (*кожаный, глиняный, деревянный, земляной*); • предназначенный для помещения чего-либо (*дровяной, платяной*); • работающий на том, что названо исходным словом (*ветряной, нефтяной*)
-аст- (-ат-)	прилагательные, называющих части тела человека или животного, внешних качеств человека, аксессуаров его внешности (*волосатый, косматый, губастый, очкастый, рогатый, скуластый*)
-ев- (-ов-), [-й-]	прилагательные со значением: • принадлежности предмета лицу или животному (*дедов, слесарев, волч[й+а], собач[й+а]*); • сделанный из чего-либо, относящийся к кому-либо, чему-либо (грушевый*, садовый*)
-еват- (-оват-)	прилагательные со значением: • отчасти напоминающий кого-либо или имеющий некоторое свойство чего-либо (*мужиковатый, плутоватый, молодцеватый*); • оттенка ослабленного качества (несколько, слегка) (*голубоватый, беловатый, сладковатый*)

Продолжение таблицы 2.3

Суффикс	Значение
-енн- (-онн-)	прилагательные со значением: • признака или свойства (*клюквенный, клятвенный, утренний, традиционный*); • подверженности действию, результата действия или характерезуемости действием (*медленный, усиленный, влюбленный*)
-енск- (-инск-)	прилагательные, обозначающие географические названия (*кубинский, пензенский*)
-еньк- (-оньк-)	уменьшительно-ласкательные прилагательные: *лёгонький, голубенький*
-ехоньк- (-оханьк-)	уменьшительно-ласкательные прилагательные: *смирнехонький, горькоханький*
-ешеньк- (-ошеньк-)	уменьшительно-ласкательные прилагательные: *скорешенький, легошенький*
-ив-	прилагательные со значением: постоянного свойства, качества, склонности к чему-нибудь, обладания каким-нибудь качеством в большой степени (*ленивый, лживый, красивый, игривый*)
-ий	*охотничий* (в притяжательных прилагательных)
-ильн-	прилагательные, которые обозначают признак, характеризующийся отношением или способностью к тому, что названо мотивирующим словом (*трясильный*)
-ин-	прилагательные, обозначающие людей и животных: (*гусиный, дядин*)
ин+ск	*Ольгин – ольгинский, сестрин – сестринский*
-ист-	прилагательные со значением: • похожий на что-то (*серебристый, бархатистый*); • обладающий чем-то в большом количестве (*голосистый, ветвистый*); • имеющий склонность к какому-нибудь действию (*задиристый, отрывистый, порывистый*)
-ит- (-овит-)	прилагательные со значением: обладающий в большей степени чем-нибудь (*именитый, ядовитый, сердитый*)
-к-	прилагательные со значением: склонный к какому-нибудь действию, такой, что часто делает что-нибудь, или такой, с которым часто что-нибудь делается (*ломкий, топкий, липкий, ковкий, цепкий*) чей-то*: рыбацкий*

Продолжение таблицы 2.3

Суффикс	Значение
-л-	прилагательные со значением: • находящийся в состоянии, которое возникло в результате действия, названного исходным словом (*гнилой, умелый, усталый*); • обладания признаком, названным в исходном слове (*светлый*)
-лив-	прилагательные, обозначающие состояние, действие, свойство, склонность к чему-нибудь или обладание каким-нибудь качеством (*молчаливый, счастливый, крикливый*)
-н- (-шн-)	прилагательные со значением: • признака или свойства, относящегося к предмету, явлению, действию, месту, времени или числу, названному исходным словом (*весенний, дальний, вчерашний, домашний, тысячный*); • подверженности какому-нибудь действию или результата какого-либо действия, которое названо исходным словом (отглагольные прилагательные) (*рваный, читаный, званый, драный*)
-ск-	прилагательные со значением чей-то (по географическому названию): *голландский*
-тельн-	прилагательные со значением: • производящий или способный произвести действие (*наблюдательный, удовлетворительный*); • являющийся объектом действия или способный им стать (*желательный, осязательный*); • предназначенный для выполнения действия (*плавательный, летательный*); • указывающий на определенную связь с действием (*избирательный. подготовительный*)
-уч- (-юч-, -яч-)	прилагательные со значением: склонный к какому-нибудь действию (*певучий, вонючий, висячий*)
-чат-	прилагательные со значением: • обладающий чем-нибудь, имеющий в большом количестве или в большой степени что-нибудь (*узорчатый, бревенчатый, бугорчатый*); • наполняющий каким-нибудь качеством, свойством то, что обозначается исходным словом (*дымчатый, дудчатый, репчатый*)
-чив-	прилагательные со значением: способный, склонный что-нибудь делать, проявлять какое-нибудь свойство (*находчивый, сговорчивый, устойчивый*)
ян+ист	*Маслянистый*

Таблица 2.4 – Суффиксы глаголов

-а- (-я)	глаголы от существительных с общим значением действия (*завтракать, козырять*)
-а-, -ка-	глаголы от междометий, звукоподражательных слов (*охать, хихикать, мяукать*)
-е-	глаголы со значением: становиться кем-нибудь, каким-нибудь, быть таким, каким называет исходное слово (*стареть, звереть, темнеть, богатеть, хорошеть*)
-и-	глаголы со значением: • совершать действие, свойственное тому, что названо в исходном слове (*батрачить, рыбачить*); • превращать в кого-либо, делать кем-либо (*калечить*); • действовать с помощью предмета (*боронить, сверлить*); • иметь место, происходить (о явлениях природы) (*морозить, моросить, порошить*); • наделения качеством, названным в исходном слове (*румянить, чернить*); • приведения в состояние, названное в исходном слове (*веселить, злить, печалить*); • удаления чего-либо, названного исходным словом, с поверхности или изнутри (*потрошить, шелушить*); • действия или признака, названного в исходном слове (*хитрить, пружинить*);
-нича-	глаголы со значением: заниматься деятельностью, обнаруживать склонность к какой-либо деятельности (*попрошайничать, обезьянничать, садовничать*)
-ну-	глаголы со значением мгновенности, моментальности (*слепнуть, ахнуть, цокнуть*)
-ова-, -ева-, -ива-, -ыва-, -ствова-	глаголы со значением: осуществлять что-либо, находиться в каком-нибудь состоянии или предаваться какой-нибудь деятельности (*торговать, тосковать, горевать, воровать, блаженствовать, бесчинствовать, безмолвствовать, навьючивать, подглядывать*)
-ся (-сь)	глаголы со значением: • лица, совершающего действие, направленное на самого себя (*мыться, одеваться*); • внутреннего состояния субъекта, настроения, переживания (*радоваться, интересоваться*); • движения, совершаемого субъектом (*кататься, биться*); • действия, постоянно свойственного субъекту (*жжется* (крапива), *клюется* (петух)); • действия, совершаемого несколькими лицами (*встречаться, ссориться*); • действия, совершаемого субъектом для себя, в своих интересах (*запасаться, укладываться*); • полноты, исчерпанности, проявления действия, удовлетворенность, истощенность субъекта действием, интенсивность захвата субъекта действием (*належаться, выспаться, разгуляться*)

Продолжение таблицы 2.4

Суффикс	Значение
-ене- (-ени-)	*леденеть, леденить*
-л – суффикс прошедшего времени	*смотреть – смотрел*
-и, -ите - суффиксы повелительного наклонения	*пиши – пишите* Здесь надо различать правописание глаголов в повелительном наклонении и в изъявительном. Сравните: *Крикните ему громче!* (повелительное наклонение) *Если вы крикнете ему громче, он услышит.* (изъявительное наклонение)

Таблица 2.5 – Суффиксы причастий

Суффикс	Значение
-ущ- (-ющ-)	*пишущий, колющий* (от глаголов 1-го спряжения)
-ащ- (-ящ-)	*лечащий, клеящий* (от глаголов 2-го спряжения)
-вш-, -ш-	*коловший, несший*
-ем-	*читаемый* (от глаголов 1-го спряжения)
-им-	*клеимый* (от глаголов 2-го спряжения)
-нн-, -енн-, -т-	*нарисованный, обиженный, помятый*

Таблица 2.6 – Суффиксы деепричастий

Суффикс	Значение
-а (-я)	деепричастия несовершенного вида: *стуч-а, грем-я*
-в, -вши, -ши	деепричастия совершенного вида: *получи-в* (от получить - глаголов с основой на гласный), *получи-вши-сь* (от получиться), *истек-ши* (от истечь - глаголов с основой в неопределенной форме на согласный)

Таблица 2.7 – Суффиксы наречий

Суффикс	Значение
-а, -о, -е	наречия со значением оценки действия (*певуче, сильно, стремительно, снова, умоляюще, растроганно, смягчающе, взволнованно*)
-и	наречия от основ прилагательных, имеющих суффикс -ск- (*дружески, логически, систематически*)
-жды	наречия от количественных числительных (*однажды, дважды*)
-учи (-ючи)	наречия от основ глаголов (*играючи, крадучись*)
-то, -либо, -нибудь	наречия, в которых при самом обобщенном указании на место, время, образ действия они остаются неясными (*где-то, где-либо, куда-нибудь, когда-то, как-нибудь*)

2.4 Примеры морфемного разбора слов разных частей речи

Рассмотрим примеры разборов слов некоторых частей речи, проводимых человеком.

Первое слово глагол ***собирался*** [12]**.**

1. нулевое окончание указывает на форму глагола прошедшего времени, ед.ч., м.р.;
2. основа – *собирался*;
3. *ся* − суффикс возвратных глаголов;
4. *л* − суффикс глаголов прошедшего времени;
5. *а* − суффикс глагольной основы;
6. *со* − приставка, имеет значение объединения
7. корень -*би*р-; возможно чередование *бир/бер/бр*.

Отобранного – причастие.

1. окончание -*ого* - указывает на форму причастия ед.ч. м.р.;
2. основа – *отобран*;
3. *нн* − суффикс страдательного причастия прош.вр.;
4. *а* − суффикс глагольной основы;
5. *ото* − приставка со значением отделения;
6. корень *бр*- возможно чередование *бр/бер/бир*.

Обижает – глагол.

1. окончание -*ет* - указывает на форму глагола наст.вр., 3 л.,ед.ч.;
2. основа − *обижай*;
3. *й* − суффикс глаголов наст. вр.;
4. *а* − суффикс глагольной основы несов.в. (обидеть);
5. корень *обиж* - возможно чередование *обиж/обид*.

Городской – прилагательное.

1. окончание –*ой*, прилагательное в форме м. р., именительного падежа, ед.ч.;
2. основа − *городск*-;
3. *ск* − словообразовательный суффикс;
4. корень − *город*-.

Подоконник – существительное.

1. нулевое окончание – так как часть речи существительное в форме м. р., именительного падежа, ед.ч;
2. основа – *подоконник*;
3. *ник* – словообразовательный суффикс;

4. *под* – словообразовательная приставка со значением – внизу, ниже низа предмета;

5. корень – *окон-*.

Весело – наречие.

1. окончания нет – так как наречия не имеют окончаний;
2. *о* – суффикс наречия со значением оценки действия;
3. корень – *весел-*.

Очень интересный разбор слова ***вынуть***. Это единственное слово в русском языке не имеющее корня! *Вы-* приставка со значением изнутри, *ну-* – суффикс глагола со значением мгновенности, моментальности, *ть-* – по одним источникам суффикс (З.А.Потиха, М.Т.Баранов, Д.Э.Розенталь) [13], по другим – окончание глаголов (П.А.Лекант, А.Н.Тихонов, А.В.Дудников, Н.С.Валгина) [13] неопределенной формы.

Таким образом, мы рассмотрели разбор слов по составу глаголов, существительного, прилагательного, наречия и причастия.

3 Математическая модель морфемного анализа слов русского языка

Чаще всего для разбора слов по составу используют следующий порядок:

1. Определить слово как часть речи;
2. У изменяемого слова найти окончание и определить его значение. Правильность выделения окончания проверить его изменением;
3. Указать основу слова;
4. Выделить корень (для этого нужно подобрать однокоренные слова) или корни в сложных словах;
5. Выделить приставки и суффиксы (если они есть). Правильность выделения морфем доказать подбором слов с другим корнем, но с этими же приставками и суффиксами.

В отличие от человека, компьютер не сможет провести аналогии между словами, он не сможет построить образы и вычленить из слова его составные части, следовательно, для таких действий ему необходим чёткий алгоритм разбора. Минимальной значимой единицей слова является морфема. Наиболее часто выделяют: приставку, корень, суффикс, окончание, основу и соединительную гласную (только для сложных слов).

Точно известно, что приставка может находиться в начале слова или ее может не быть вообще. С помощью приставки, образование слов обычно идет в пределах одной и той же части речи. Аналогично с окончанием, оно всегда в конце слова, однако его может и не быть (нет окончаний у служебных слов, наречий, неизменяемых существительных и прилагательных). Окончание изменяемая часть, которая служит для связи слов и выражает значения рода, числа, падежа, лица [8]. Для того чтобы выделить окончание необходимо знать к какой части речи относится слово и как изменяется в результате склонения. Проанализировав результаты изменившейся в слове части, можно выделить ее окончание.

Все что находится между приставкой и окончанием, будет либо корнем, либо суффиксом, либо соединительной гласной. Корень обязательная часть слова, он может стоять за приставкой, либо стоять в начале слова. Существуют слова, состоящие только из корня. Есть слова состоящие из нескольких корней, некоторые из них могут соединяться с помощью гласных *–о-, -е-* (соединительных гласных).

После корня следует суффикс, которого так же может и не быть. С помощью суффиксов, в отличии от приставок, образуются новые слова, которые могут принадлежать различным частям речи. Именно поэтому можно выделить суффиксы свойственные той или иной части речи.

Из приведенного выше, можно сделать вывод, что автоматизированный разбор слова по составу невозможен без определения части речи, к которой он относится.

Таким образом, наиболее логично изменить классический порядок морфемного разбора слова на следующий:

1 Определить часть речи слова;

2 Определить окончание (если оно существует);

3 Определить возможные приставки;

4 Определить корень (корни и соединительные гласные);

5 Определить возможные суффиксы;

6 Выделить основу.

Таким образом, математическое описание задачи следующее:

Входные данные:

Slovar =$\{w_a\}$,

где *Slovar* - множество всех слов русского языка,

w_a − слово русского языка;

Выходные данные:

$w_a = \{osnova, < prefix >, < root >, < suffix >, okonchanie\}$,

где $osnova, prefix, root, suffix, okonchanie$ – морфы слова.

Алгоритм:

1 Пусть *Slovar* – множество всех слов русского языка, тогда $\forall w_a \exists pOfs\ (P(w_a, pOfs)\ \&\ w_a \in Slovar)$,

где $P(w_a, pOfs)$ = *Слово* w_a *является часть речи* $pOfs$

2 Если $\exists w_a\ (\overline{E(w_a)})$, то $w_a = \{osnova\}$, иначе $w_a = \{osnova, okonchanie\}$,

где E(w_a) = *слово имеет окончание.*

3 Если $\exists w_a \exists prefix\ \ (Pr(w_a, prefix\ ))$, то $w_a = \{osnova, < prefix > , okonchanie\}$, иначе $w_a = \{osnova, okonchanie\}$,

где $Pr(w_a, prefix\ )$ = слово имеет префикс $prefix$

4 Если $\forall w_a\ \exists root\ \ \left(w_a <> \textit{исключение} \Rightarrow R(w_a, root\ )\right)$, то $w_a = \{osnova, < prefix >, < root >, okonchanie\}$,

где $R(w_a, root\ )$ = слово имеет корень $root$

5 Если $\exists w_a \exists\ suffix\ \ (P(w_a, pOfs)\ \&\ Sf(w_a, suffix\ , pOfs\ ))$, то $w_a = \{osnova, < prefix >, < root >, < suffix >, okonchanie\}$, иначе $w_a = \{osnova, < prefix > , < root >, okonchanie\}$,

где $Sf(w_a, suffix\ , pOfs\ )$ = слово имеет суффикс $suffix$, который относится к части речи $pOfs$

4 Функциональная модель системы и её структура

В данном разделе будут представлены функциональная модель системы и ее структурная схема. Основной целью функциональной модели является определение минимального набора функций, необходимых для разработки системы морфемного анализа слов русского языка. Модель будет рассматриваться точки зрения пользователя системы. Структурная схема дает представление о взаимной связи отдельных элементов описываемой системы.

На рисунках 4.1-4.5 представлена функциональная модель «to be», созданная для проектирования системы морфемного анализа слов русского языка. Функциональное моделирование выполняется средствами программы ErWin Process Modeler [14] по стандарту IDEF0 [15].

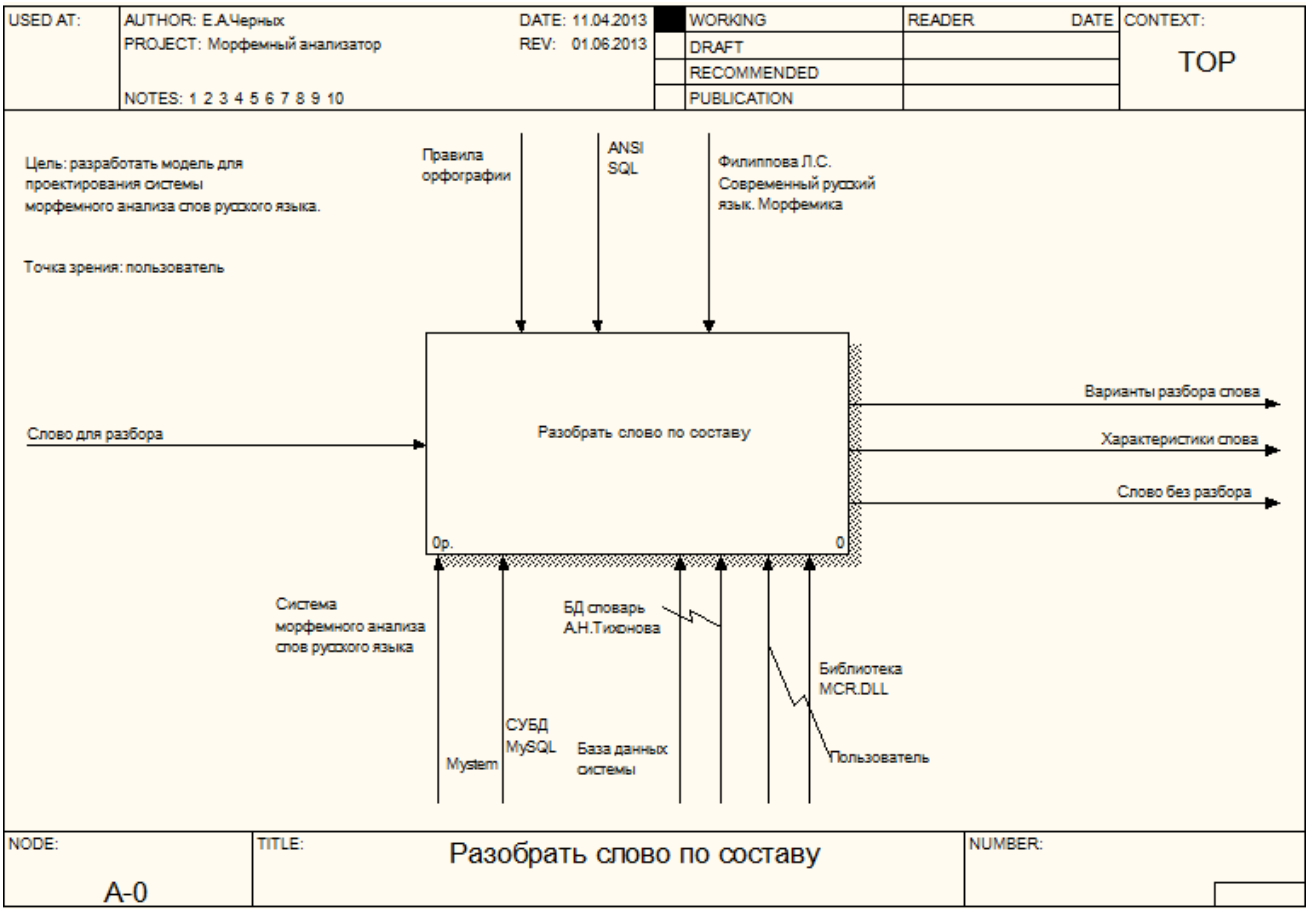


Рисунок 4.1 – Контекстная диаграмма «Разобрать слово по составу»

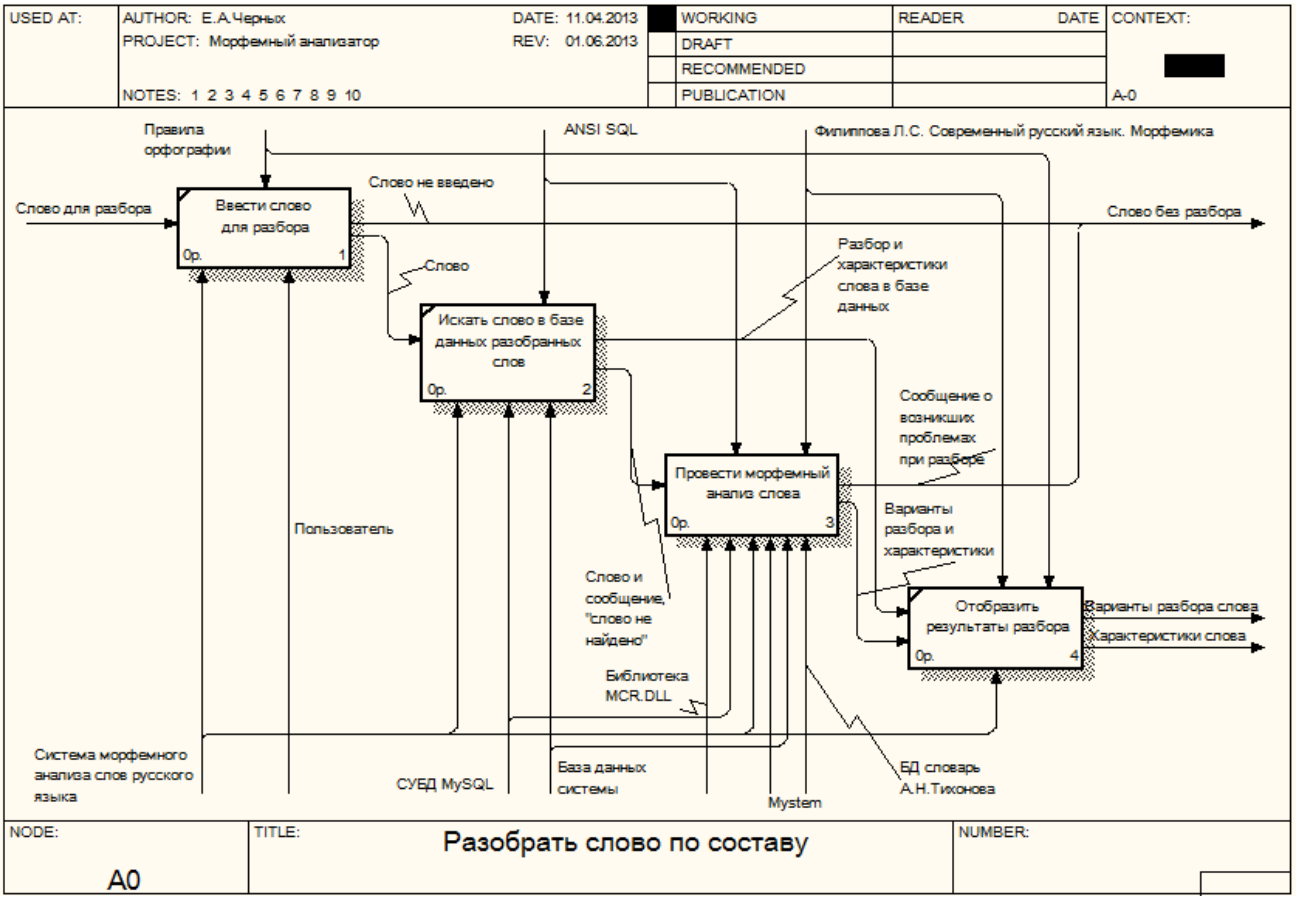


Рисунок 4.2 – Дочерняя диаграмма A0 «Разобрать слово по составу»

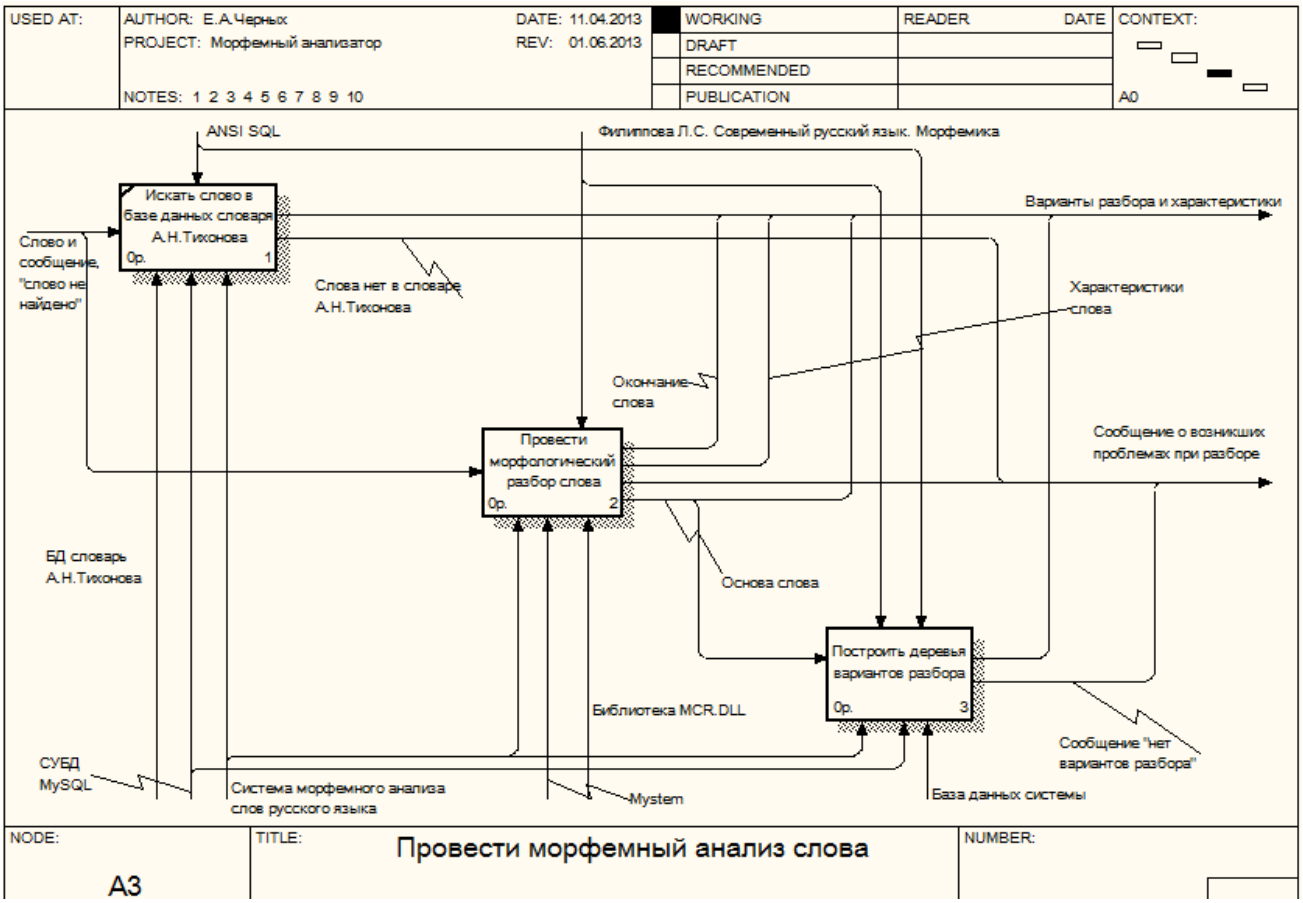


Рисунок 4.3 – Дочерняя диаграмма A3 «Провести морфемный анализ слова»

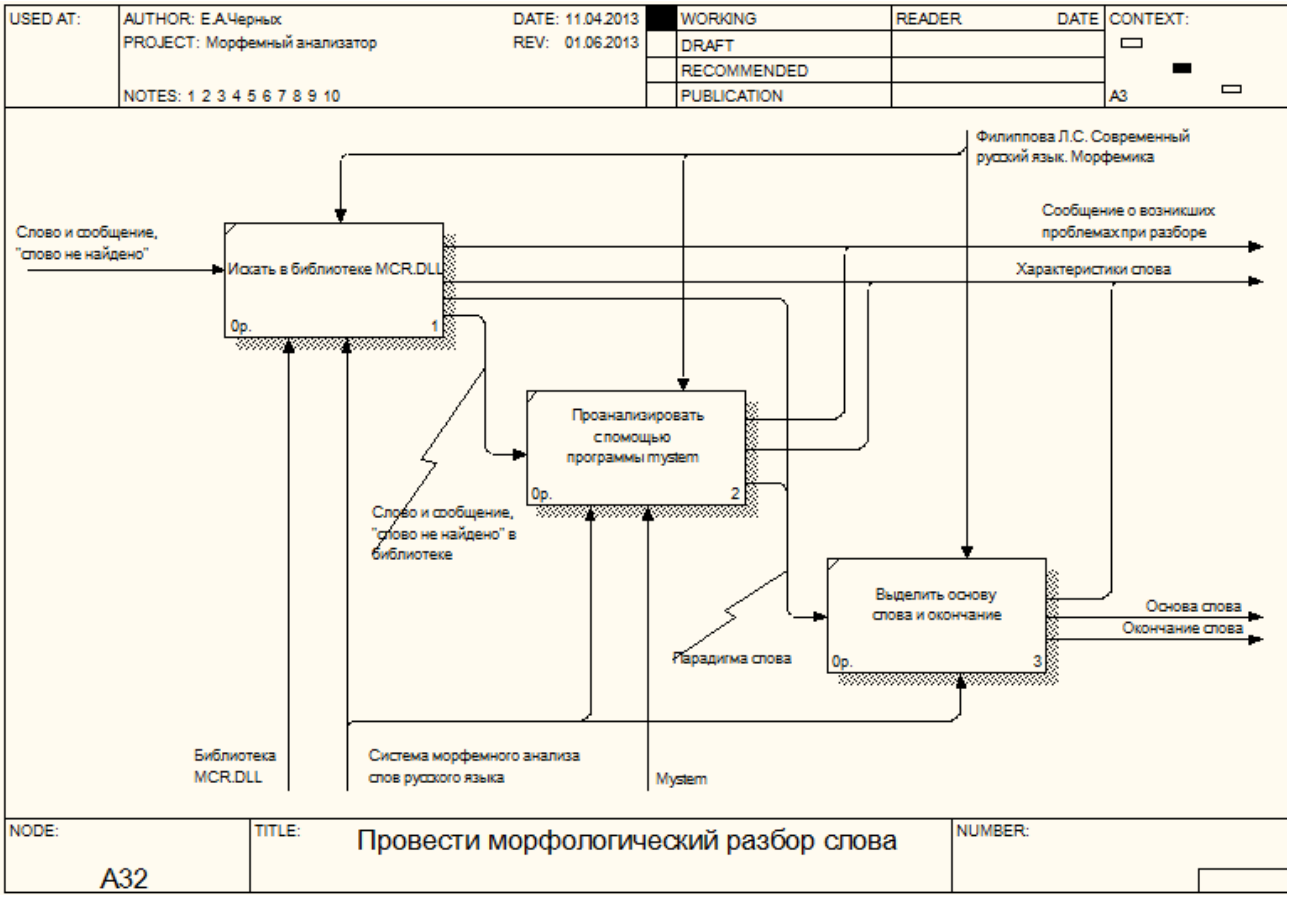


Рисунок 4.4 – Дочерняя диаграмма A32 «Провести морфологический разбор слова»

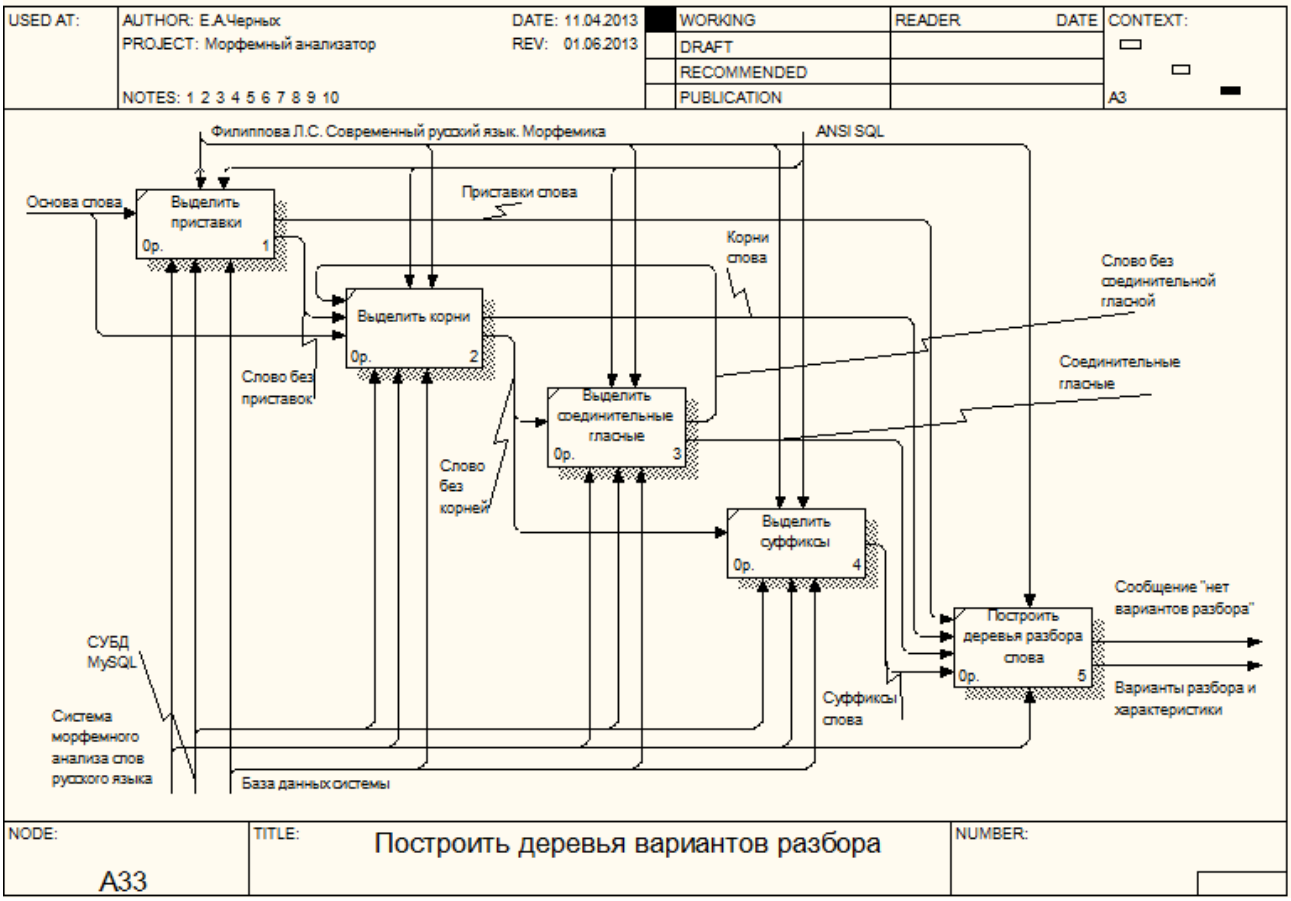


Рисунок 4.5 – Дочерняя диаграмма A33 «Построить деревья вариантов разбора»

На основании функциональных диаграмм можно сделать вывод, что при разработке системы необходимо создать три основных модуля: поиск готовых вариантов разбора в базе данных системы, морфемный анализ слова и представление результатов морфемного разбора. Морфемный анализ в свою очередь состоит из таких функций, как поиск вариантов разбора в морфемно-орфографическом словаре А.Н.Тихонова [4], морфологический анализ слова для выделения основы слова и окончания и морфемный разбор слова с сокращением вариантов разбора слов.

Функция сокращения вариантов разбора слов заложена в функциональном блоке A335, более подробно алгоритм сокращения вариантов разбора представлен в главе 5.2 «Разработка алгоритма сокращения вариантов морфемного разбора». Представлению результатов так же выделена отдельная глава отчета (глава 5.3 «Разработка подсистемы визуализации результатов морфемного разбора»).

Определив минимальный набор функций системы, можно разработать ее структурную схему (рисунок 4.6). Модули системы выделены в соответствии с их функциональным назначением. Основная цель программы автоматизация морфемного анализа слов, для ее достижения достаточно разработать модули выделенные как Основные. Для дальнейшей работы над системой и накопления информации о морфемном анализе, необходимо дополнить систему дополнительными модулями, они обозначены как Модули расширения функций. Более подробно об их разработке рассказано в главе 5.4 данной пояснительной записки.

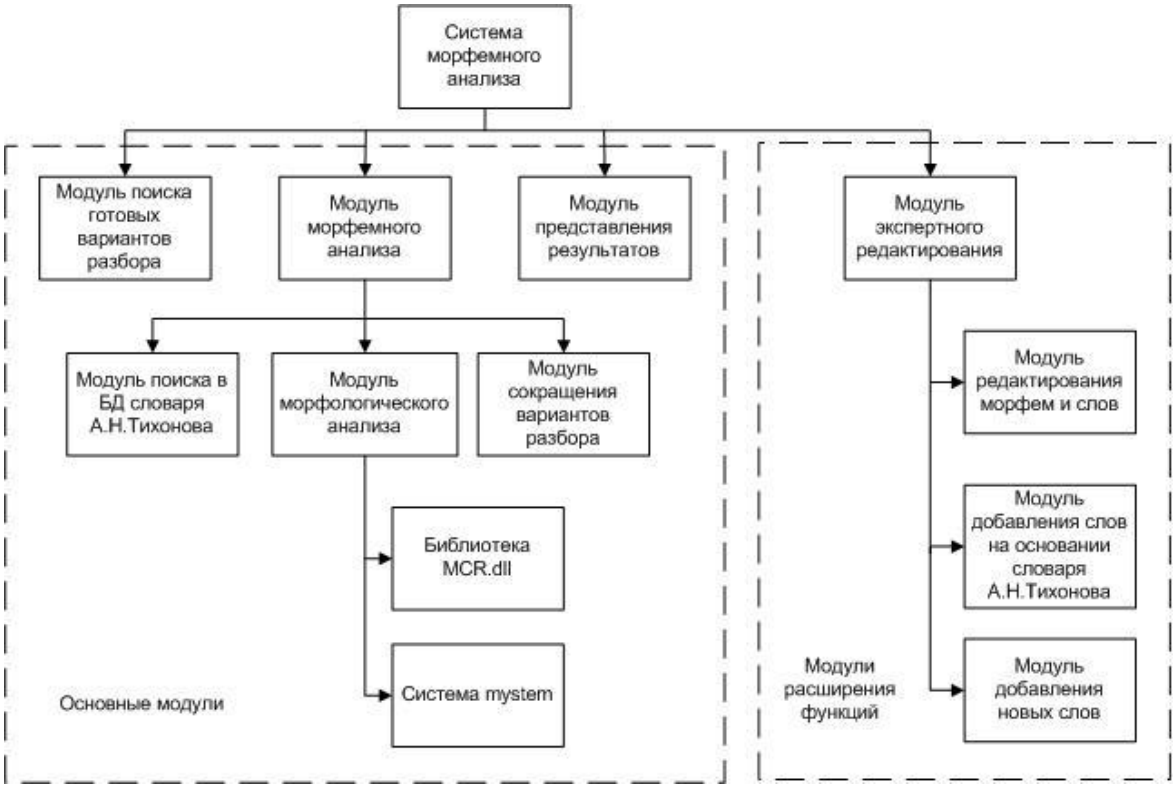


Рисунок 4.6 – Структурная схема

5 Проектирование программной системы

5.1 Разработка подсистемы морфологического анализа слов русского языка на основе библиотеки MCR.dll и программы mystem

Первым пунктом в установленном нами разборе слова по составу является определение части речи, которой принадлежит слово. В зависимости от части речи, можно определить, является ли слово изменяемым, возможно ли выделение окончания и, в какой-то степени, можем ограничить количество возможных вариантов суффиксов данного слова.

Для определения части речи было принято использовать библиотеку MCR.dll [16], созданную на основе грамматического словаря русского языка А.А.Зализняка. Библиотека позволяет определить часть речи слова и некоторые другие его характеристики. В библиотеке часть речи слова хранится как параметр постоянной грамматической характеристики слова.

Морфология (от греч. morphe «форма», logos « учение, наука») – раздел грамматики, изучающий структуру слова и выражение грамматических значений в слове [16].

Следующим пунктом разбора слова по составу является определение окончания слова. Окончанием слова называется его изменяемая часть, которая служит для связи слов и выражает значения рода, числа, падежа, лица [7]. Таким образом, у наречий, служебных слов, неизменяемых существительных и прилагательных, а так же у деепричастий и глаголов в форме инфинитива, выделить окончание невозможно. Т.е. уже на этапе определения части речи слова, можно сделать выводы о его морфемном составе.

У изменяемых частей речи окончания выражают следующие грамматические значения [8]:

- род, число, падеж – у существительных, прилагательных, причастий, некоторых местоимений, некоторых числительных;
- лиц и число – у глаголов в настоящем времени;
- род и число – у глаголов в прошедшем времени, у кратких прилагательных.

Для того чтобы выделить окончание, необходимо изменять слово по роду, числу, лицу, падежу, в зависимости от части речи. Сравнив полученные значения слова, можно выделить окончание и основу. Основа слова – это часть слова без окончания. Основа слова может быть равна корню. Кроме корня основа может включать приставки, суффиксы, соединительные гласные. Для выделения парадигмы слова (парадигма включает значения слова при изменении рода, числа, лица, падежа), так же будет использоваться библиотека MCR.dll.

Сравнив парадигму и выделив в ней неизменяемую часть, система выделит окончание и основу слова.

Значение характеристик парадигмы хранится как параметр переменных грамматических характеристик слова.

В целом структура парадигмы слова библиотеки MCR.dll выглядит следующим образом (рисунок 5.1):

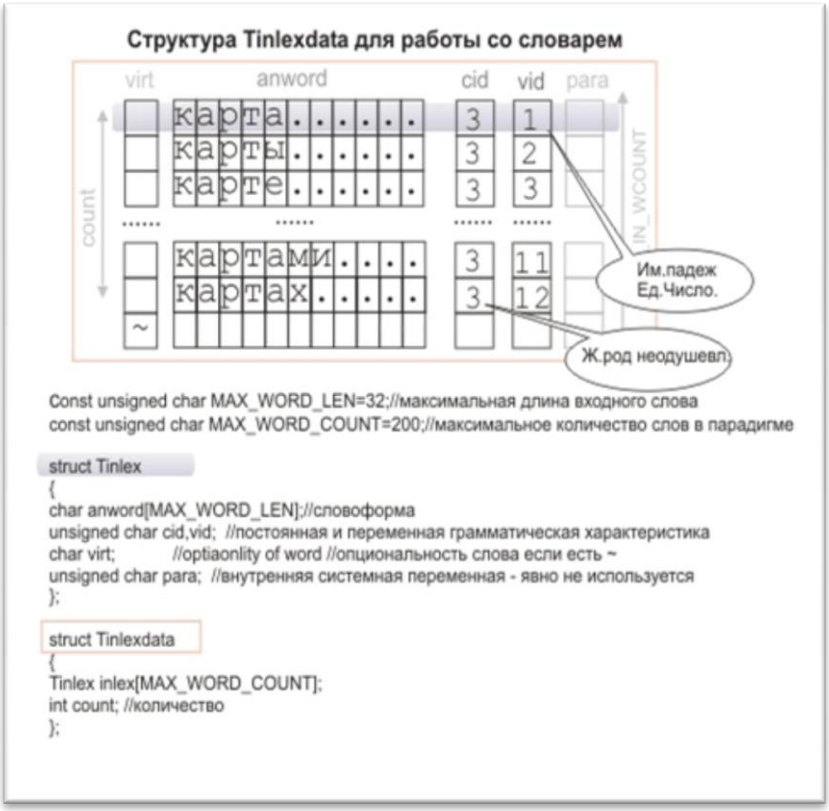


Рисунок 5.1 – Структура парадигмы для работы со словарем MCR.dll

Анализируя значения Cid и Vid, получаем парадигму слова и, соответственно, окончание и основу слова. Значения, которые могут принимать параметр Cid и Vid приведены в Приложении В «Таблицы кодирования переменных и постоянных грамматических характеристик» данного отчета.

Таким образом, цель морфологического анализа состоит в получении основы и окончания слова со значением грамматической категории (например, часть речи, род, число, падеж) для каждой из словоформ слова, поступивших на вход системы.

5.2 Разработка алгоритма сокращения вариантов морфемного разбора

В процессе разработки системы, встала задача сокращения вариантов морфемного разбора. Система строит деревья, состоящие из всех возможных вариантов сочетаний морфем. Часть веток стает ошибочными. То есть встречаются случаи, когда сумма всех узлов ветки не дает в результате основу слова, введенного для разбора. Либо разбор заканчивается приставкой или соединительной гласной, что противоречит правилам морфемного разбора. Для устранения таких веток был разработан алгоритм сокращения вариантов разбора слов. Графическое изображение алгоритма представлено на рисунках 5.2.

Описание алгоритма:

1. Начиная с корневого узла дерева, поочередно проходим по всем узлам и ищем те, которые не имеют узлов-потомков;
2. Если узел не имеет потомков, то запоминаем его и ищем следующий «тупиковый» узел;
3. Если тупиковый узел найден, то ветку с предыдущим узлом отправляем на проверку;
4. Повторяем пункты 2-3, пока не будут пройдены все узлы дерева;
5. Так как последняя ветка всегда остается не проверенной в конце ее всегда отправляем на проверку.

Необходимо отправлять ветку на проверку только при нахождении следующей тупиковой, так как в противном случае будут утеряны связи между узлами, и нельзя будет обойти все узлы дерева.

Алгоритм содержит подпрограмму «Проверить ветку с узлом 2». Детализация подпрограммы дана на рисунке 5.3. Подпрограмма проверяет, не заканчивается ли разбор слова префиксом или соединительной гласной, иначе удаляет такие варианты разбора. Суммирует значения узлов ветки до корня и, если сумма не совпадает со значением введенного для разбора основания, удаляет узлы ветки, которые не имеют других потомков.

В результате использования алгоритма из дерева разбора слова удаляются все ветки, которые не удовлетворяют правилам морфемного анализа слов русского языка.

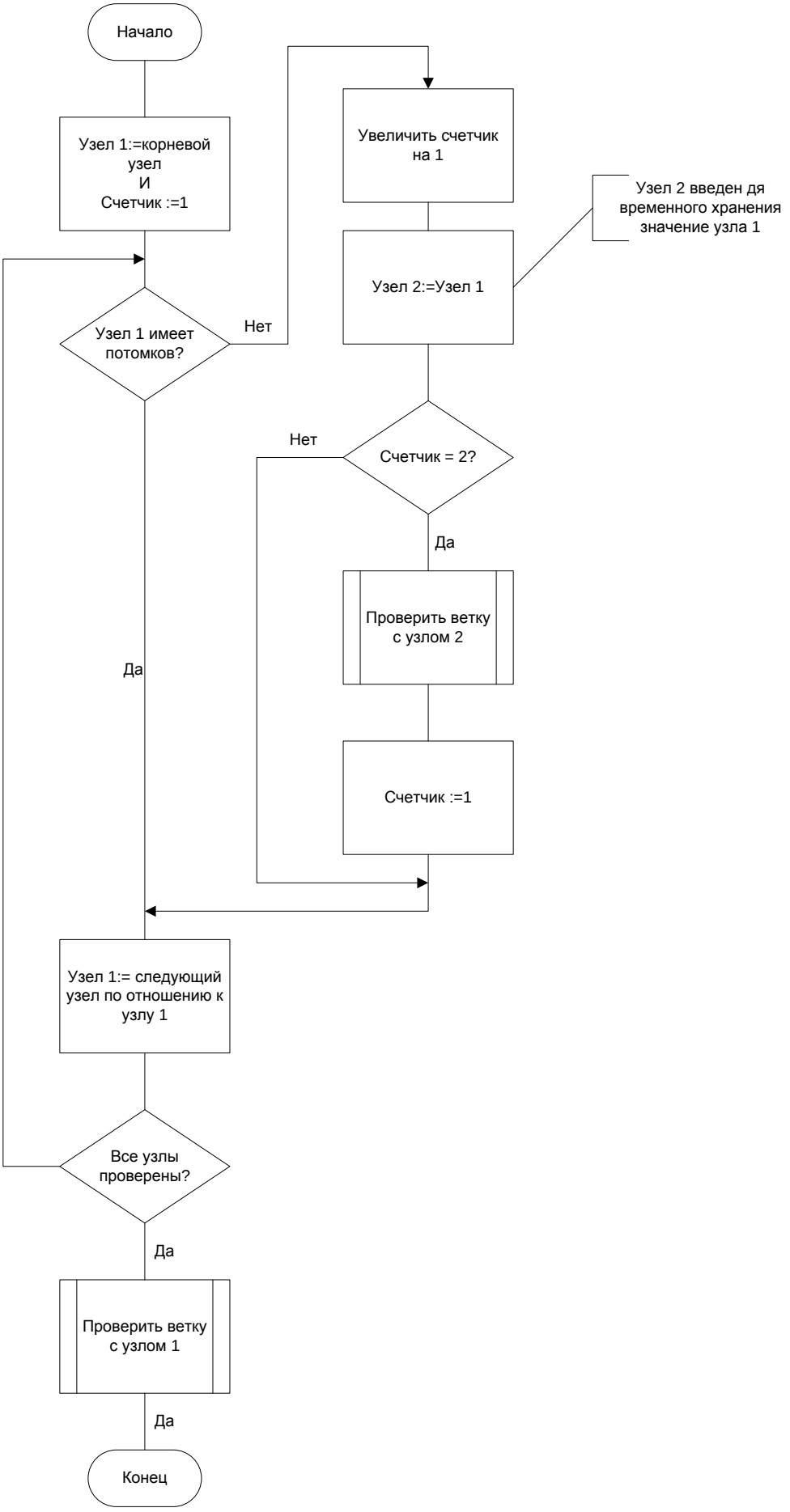


Рисунок 5.2 – Схема алгоритма сокращения результатов морфемного анализа

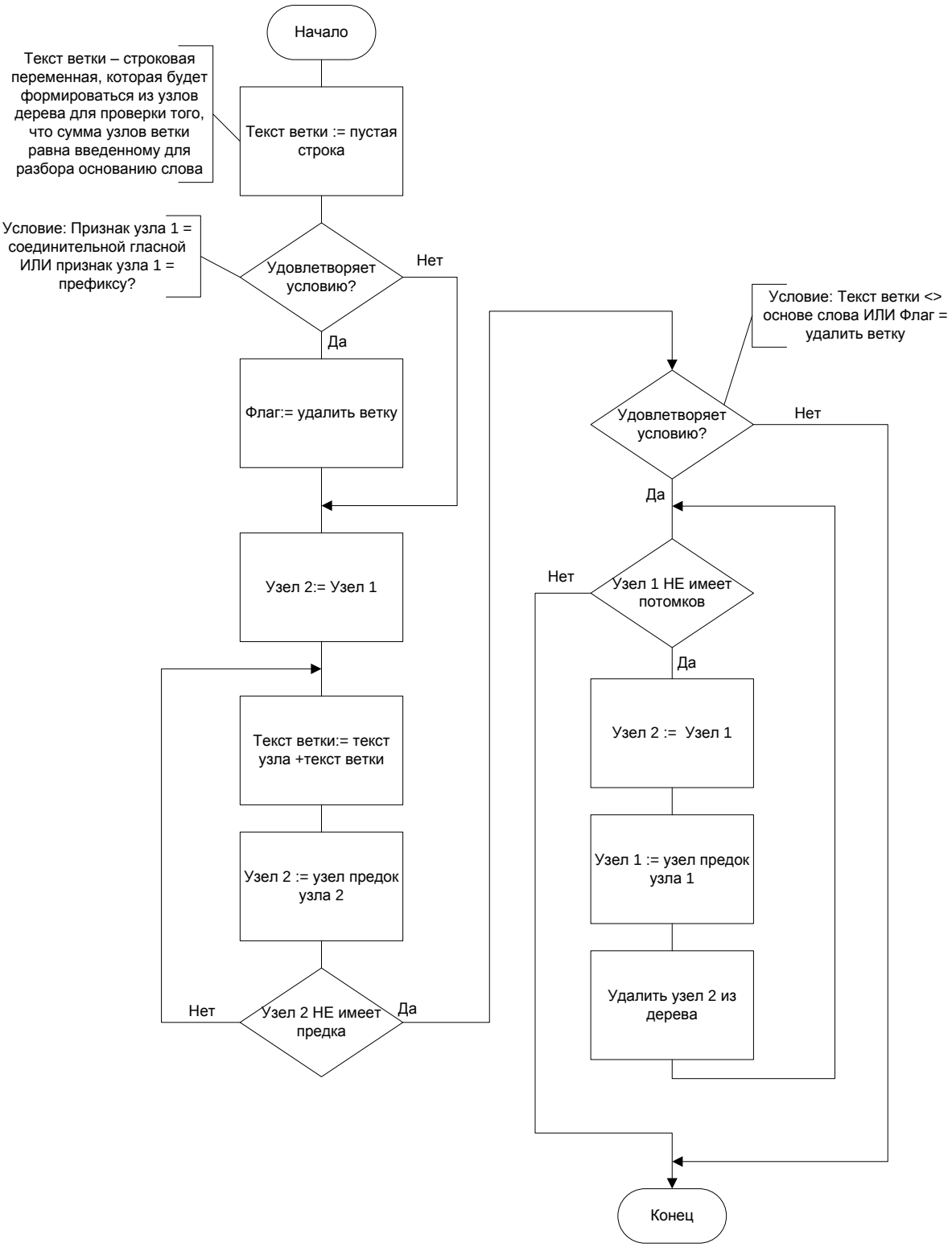


Рисунок 5.3 – Схема процедуры «Проверить ветку с узлом 2»

5.3 Разработка подсистемы визуализации результатов морфемного разбора

Визуализация результатов заключается в представлении разбора слова в привычном для пользователя графическом виде. На рисунке 5.4 представлены изображения вариантов разбора слов по составу.

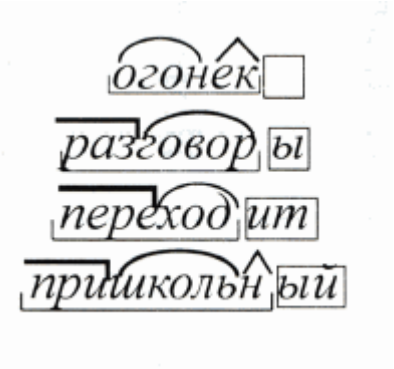


Рисунок 5.4 – Графическое представление разбора слов по составу

Знаком дуги обозначаются корни слова, значком крышечки – суффиксы, окончания выделены четырехугольником, приставки – двумя прямыми, разной длинны расположенные под прямым углом друг к другу, основание слова подчеркивается линией снизу с небольшими черточками с обоих концов, либо без них.

Для визуализации полученных в системе деревьев разбора слова был разработан алгоритм графического представления результатов разбора. Описание алгоритма представлено на схеме (рисунок 5.5).

Графическое изображение разбора слова создается возможностями Delphi 7 при помощи класса Canvas. Под операциями Вывести слово, Выделить окончание, Выделить основу, Выделить префикс, Выделить суффикс, Выделить корень понимаются процедуры использующие методы Canvas для вывода графических примитивов, соответствующих изображениям морфем. Вычисление положения курсора вывода так же осуществляется из значений заданной области для вывода изображения разбора.

Таким образом, была решена задача визуализации результатов разбора для удобства работы с системой пользователя.

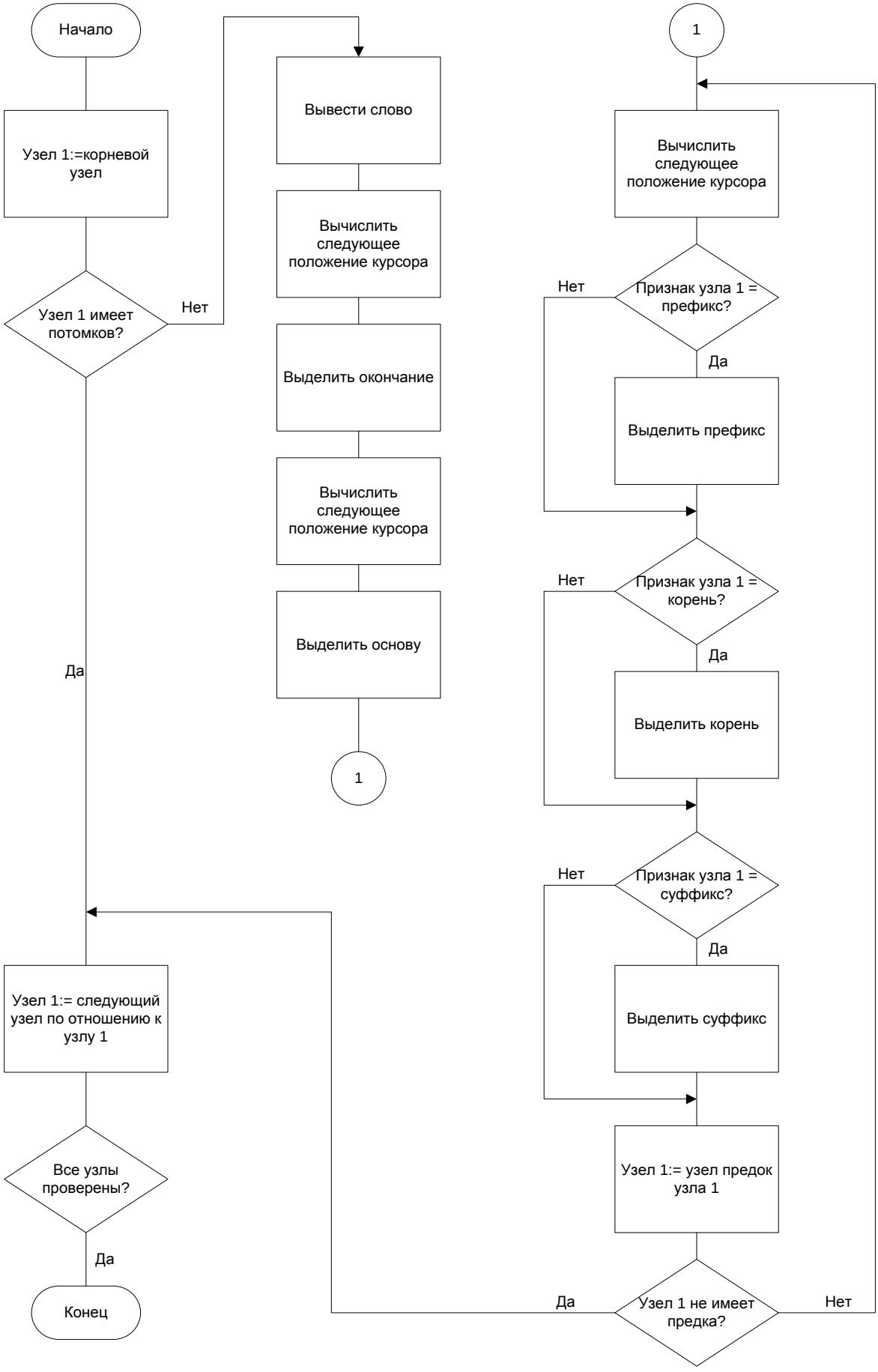


Рисунок 5.5 – Алгоритм визуализации результатов морфемного анализа

5.4 Разработка подсистемы экспертного редактирования базы данных системы

В разрабатываемой системе большое внимание уделено БД, хранящей морфемы русского языка и готовые варианты разборов. В связи с тем, что русский язык живой и развивающийся язык, в нем постоянно происходят какие-либо изменения, которые для повышения степени эффективности системы, необходимо отражать в БД. Для этого было предусмотрено разработать расширение функций системы для ряда числа пользователей (далее экспертов), которые позволят редактировать БД. Под экспертами будем подразумевать людей, профессионально занимающихся морфемным анализом русского языка, либо разработчиков системы, пополняющих ее.

На рисунках 5.6-5.10 представлена функциональная модель «to be», созданная для системы редактирования базы данных морфемного анализа слов русского языка. Функциональное моделирование выполняется средствами программы ErWin [14].

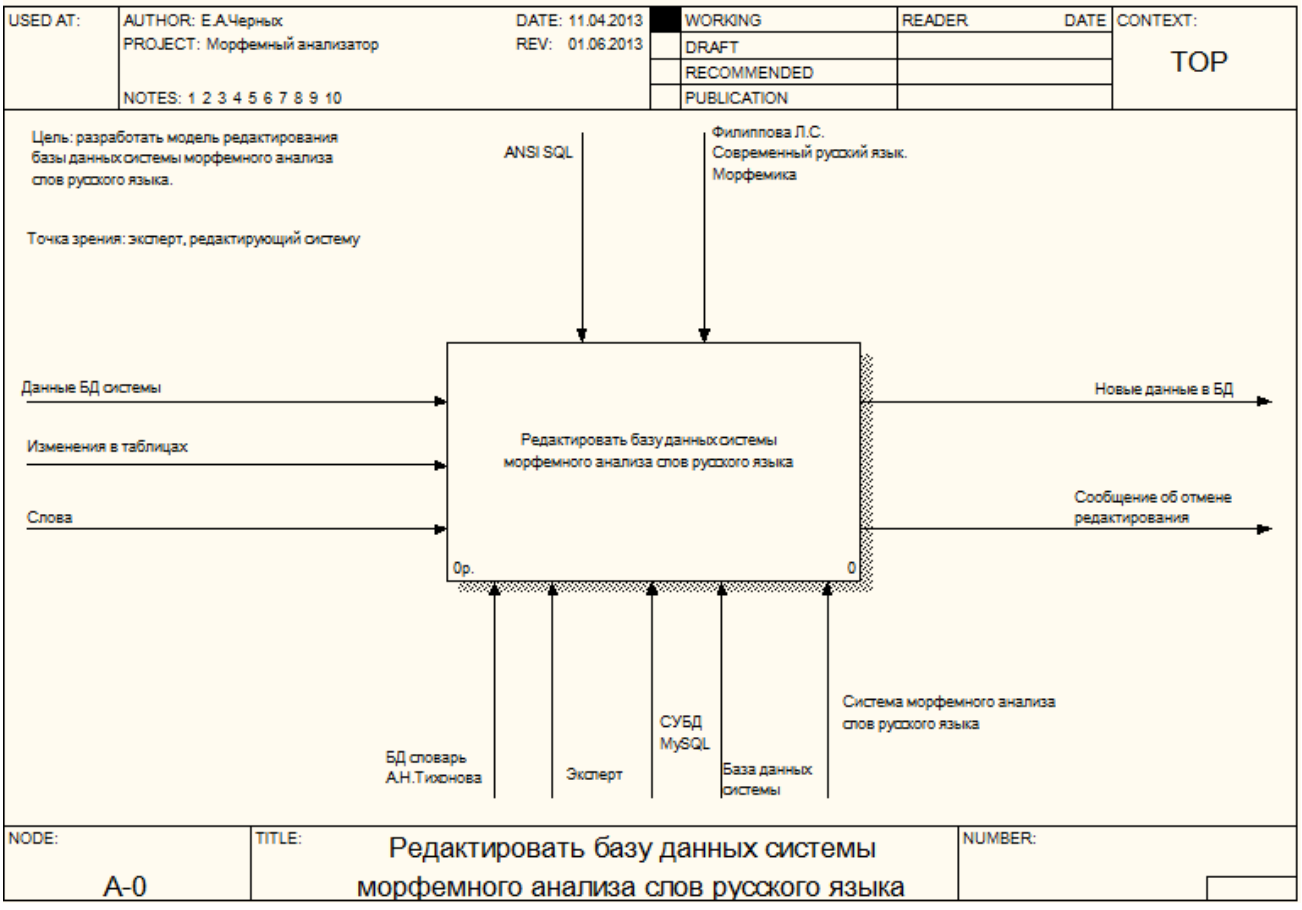


Рисунок 5.6 – Контекстная диаграмма «Редактировать базу данных системы морфемного анализа слов русского языка»

На основании функциональной модели выделяются три основных модуля: редактирование морфем, добавление новых слов как на основании словаря А.Н.Тихонова, так и отсутствующих в словаре слов и модуль выбора правильного варианта разбора. Под выбором правильного варианта разбора подразумевается выбор варианта из представленных системой вариантов разбора.

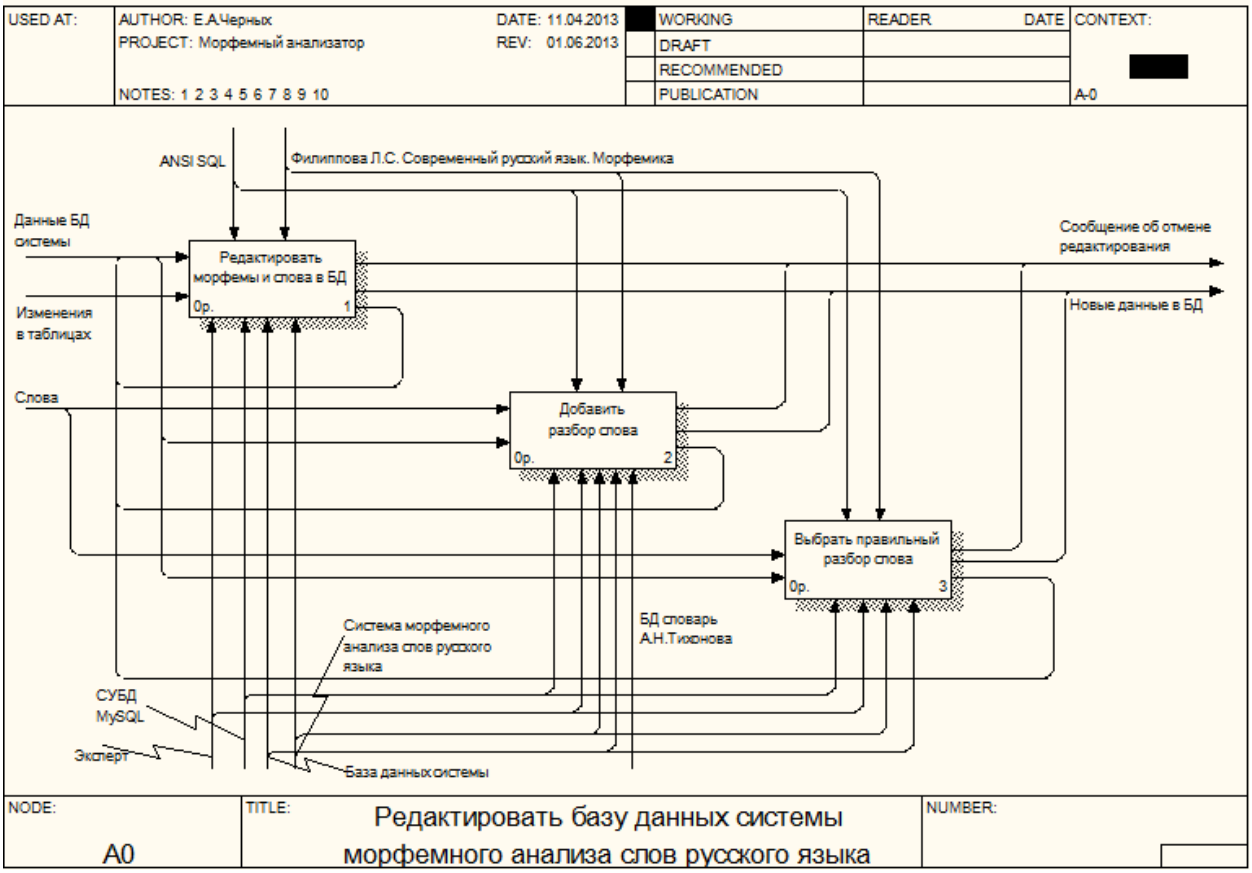


Рисунок 5.7 – Дочерняя диаграмма A0 «Редактировать базу данных системы морфемного анализа слов русского языка»

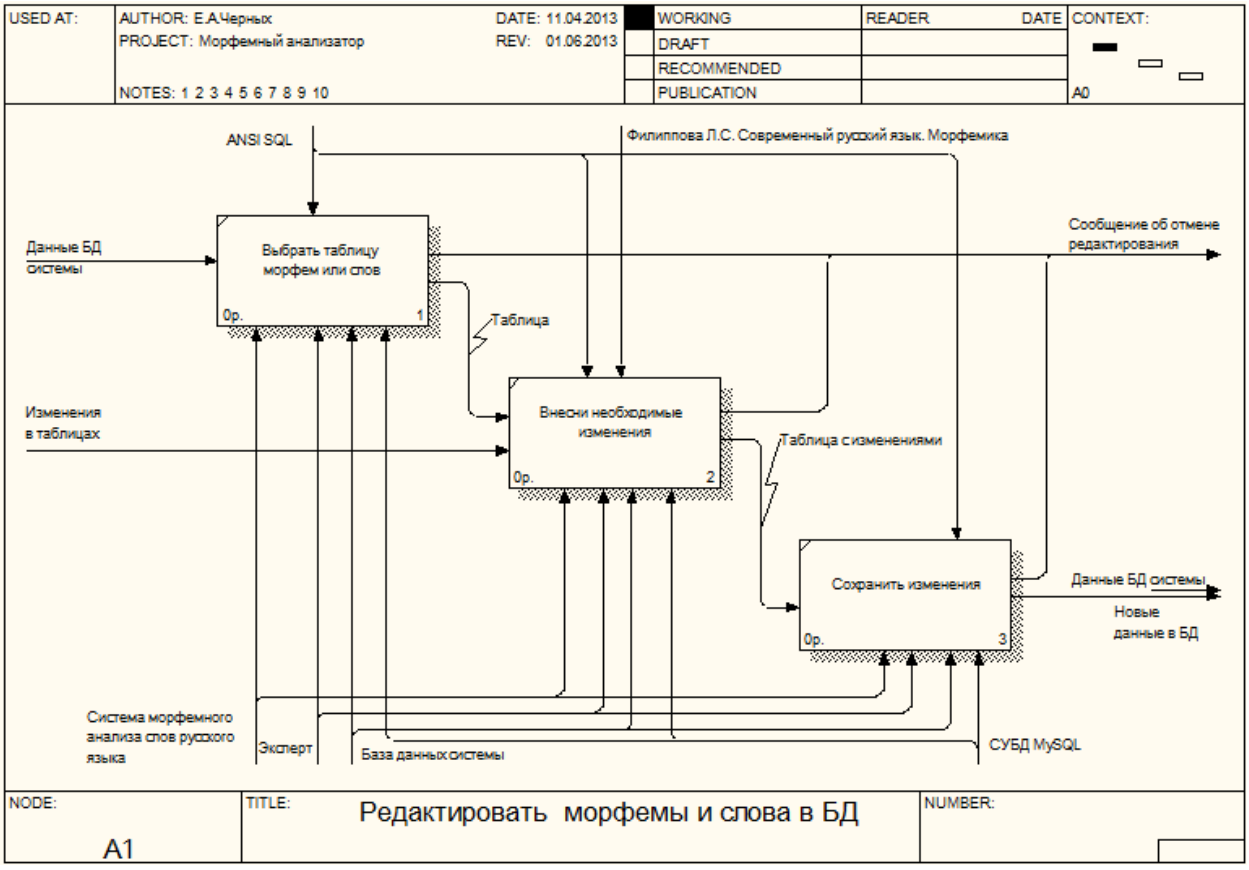


Рисунок 5.8 – Контекстная диаграмма А1 «Редактировать морфемы и слова БД»

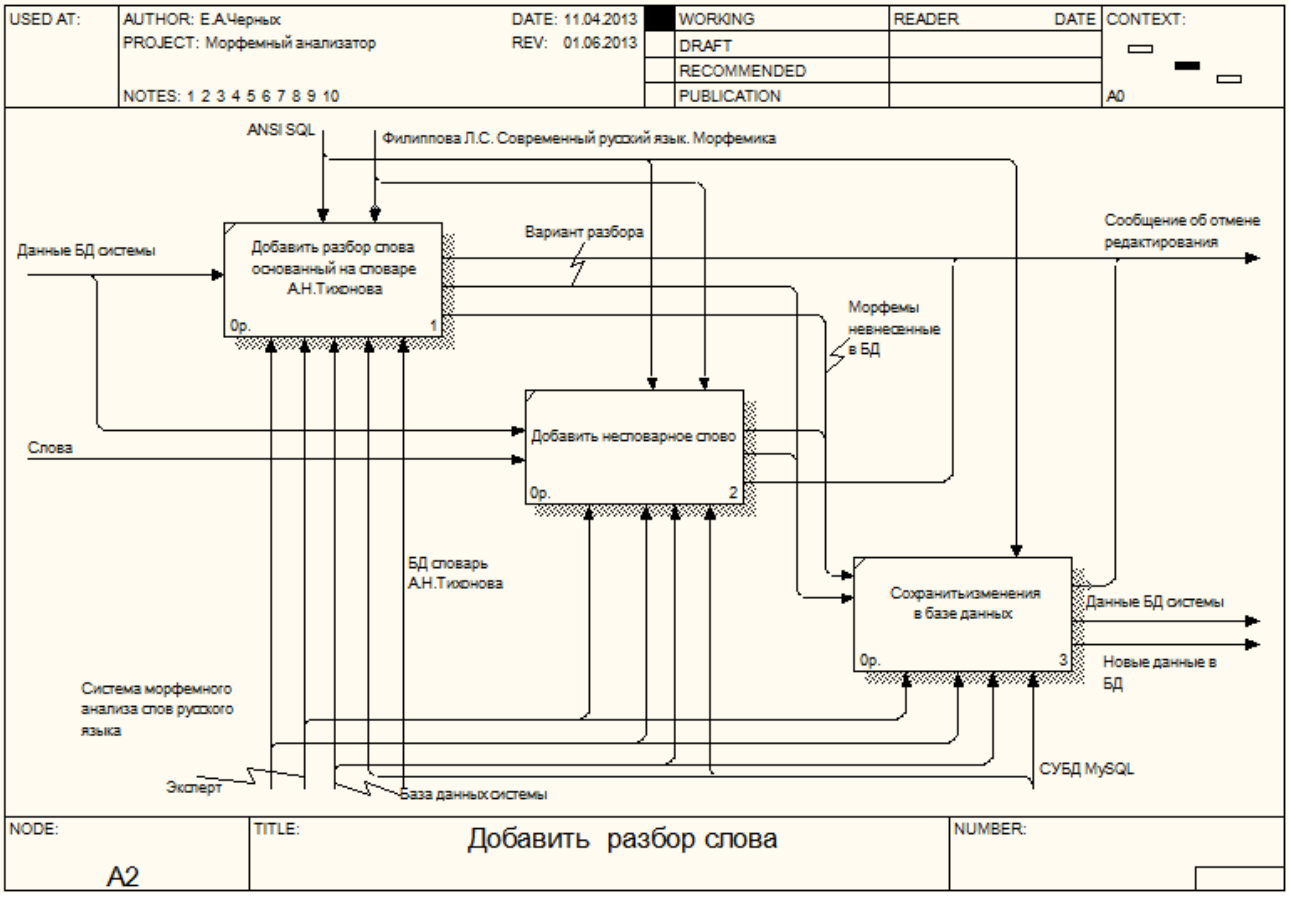


Рисунок 5.9 – Дочерняя диаграмма А2 «Добавить разбор слова»

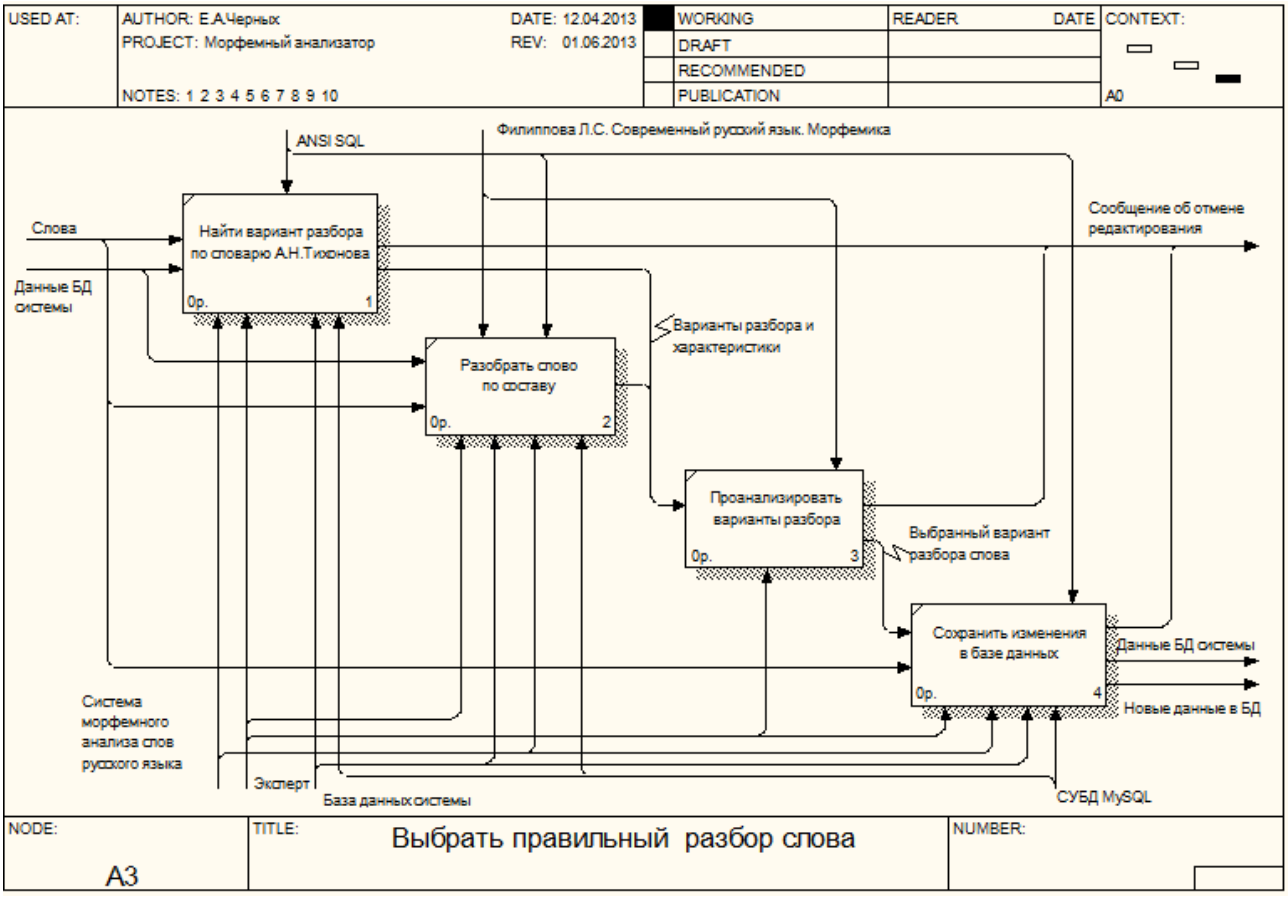


Рисунок 5.10 – Дочерняя диаграмма A3 «Выбрать правильный разбор слова»

Для проектирования подсистемы было проанализировано направлений потоков данных в ней, разработаны диаграммы DFD (рисунок 5.11).

Хранилище «Словарь А.Н.Тихонова» является электронной версией этого словаря, его редактирование и изменение не возможно и используется в системе как справочный материал. Таблица вариантов разбора и таблица морфем являются таблицами базы данных системы. Они используется в морфемном анализе слов, производимом системой. Данные этих таблиц доступны для редактирования при работе с ними эксперта.

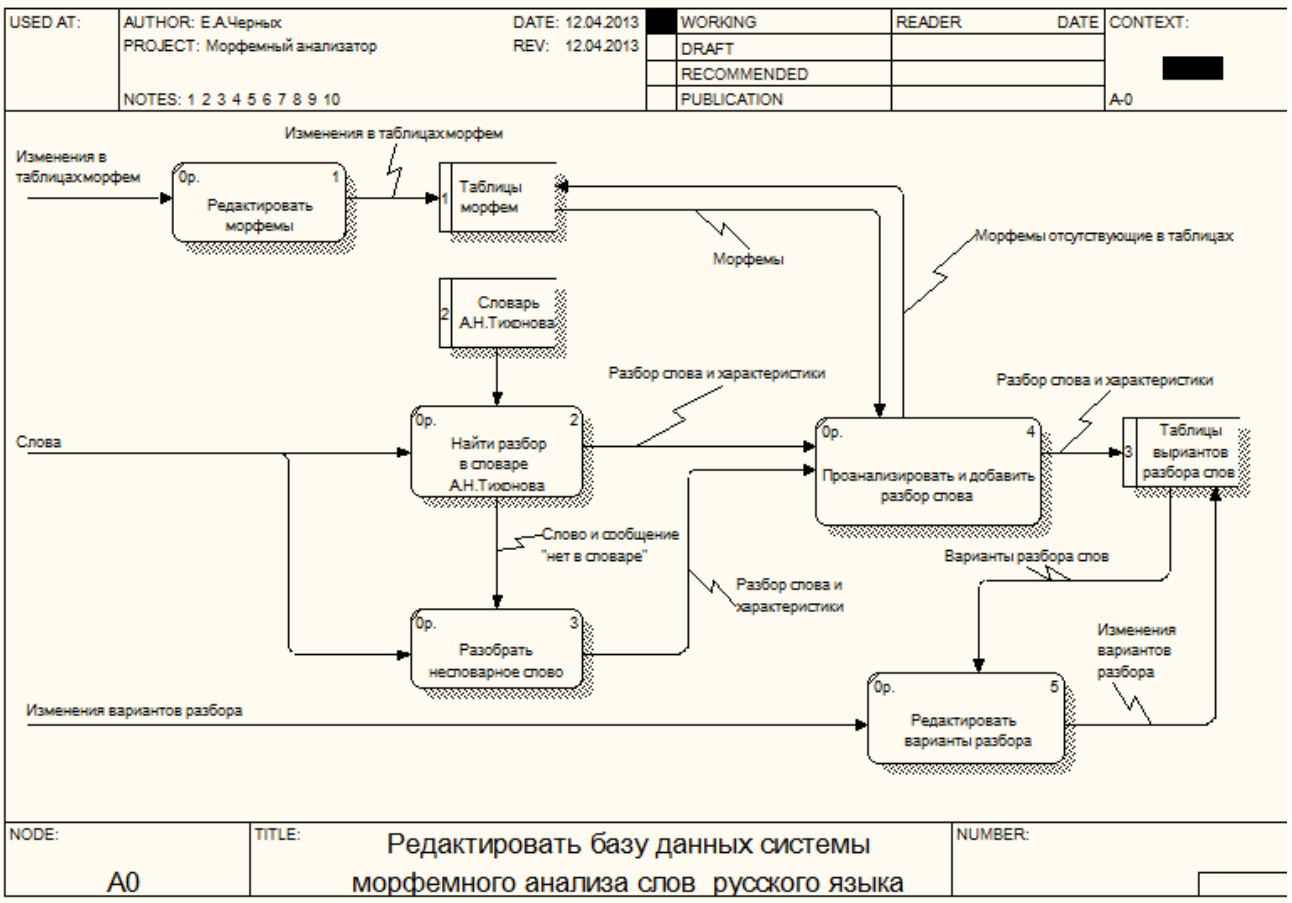


Рисунок 5.11 – Диаграмма DFD модуль системы «Редактировать базу данных системы»

6 Проектные решения по системе

6.1 Выбор и обоснование программных и технических средств

Для разработки программного обеспечения данной дипломной работы выбрана среда разработки Borland Delphi 7 [18] и реляционная СУБД MySQL [19].

Требования, которым должны удовлетворять программные средства следующие:

– обширное количество документации и простота в изучении;

– удобство в работе; гибкий, удобный, современный интерфейс работы с программами;

– простота в установке и использовании.

Delphi − это среда разработки, в которой в качестве языка программирования используется язык Delphi [20].

В среде разработки Borland Delphi 7 возможно создавать различные программы: от простейших однооконных приложений до программ управления распределенными базами. В состав пакета включены разнообразные утилиты, обеспечивающие работу с базами данных, XML-документами, создание справочной системы, решение других задач. Существует большое количество пакетов расширения возможностей Delphi 7.

Delphi 7 работает в среде операционных систем от Windows 98 до Windows 7. Стабильно работает на компьютерах стандартной конфигурации (процессор Pentium или Celeron с тактовой частотой не ниже 166 МГц, оперативной памяти - 128 - 256 Мбайт).

Язык Delphi, близок к языку Object Pascal, который является одним из базовых языков при изучении ООП.

MySQL, наиболее популярная система управления базами данных (СУБД) с открытым исходным кодом [21]. Сервер MySQL работает в клиент-серверных и встроенных системах. Сервер баз данных MySQL быстрый, надежный и простой в эксплуатации сервер.

Основные достоинства MySQL [22]:

- Записи фиксированной и переменной длины;
- ODBC драйвер в комплекте с СУБД;
- Гибкая система привилегий и паролей;
- До 16 ключей в таблице. Каждый ключ может иметь до 15 полей;
- Поддержка ключевых полей и специальных полей в операторе CREATE;
- Поддержка чисел длинной от 1 до 4 байт (ints, float, double, fixed), строк переменной длины и меток времени;

- Все операции работы со строками не обращают внимания на регистр символов в обрабатываемых строках;
- Псевдонимы применимы как к таблицам, так и к отдельным колонкам в таблице;
- Все поля имеют значение по умолчанию. INSERT можно использовать на любом подмножестве полей;
- Легкость управления таблицей, включая добавление и удаление ключей и полей.

Таким образом, связка программных продуктов Delphi 7 и MySQL удовлетворяют поставленным требования. Легко взаимодействуют друг с другом и позволяют решить задачи, поставленные для разработки системы.

6.2 Логическая модель базы данных

Для построения логической модели базы данных системы, на первом этапе определим её сущности, атрибуты и тип данных. В базе данных будет храниться информация о морфемах, т.е. необходимо ввести описание сущностей корень, префикс, окончание, соединительная гласная, суффикс. Описания атрибутов данных сущностей приведены в таблицах 6.1-6.5, соответственно.

Таблица 6.1 − Атрибуты сущности «корень»

Атрибут	Описание	Тип данных
Номер корня	Уникальный идентификатор корня в пределах таблицы	Number
Корень	Значение корня русского языка, уникальное во всей таблице	String
Комментарий к корню	Особенности использования, написания и других характеристик корня	String

Таблица 6.2 – Атрибуты сущности «префикс»

Атрибут	Описание	Тип данных
Номер префикса	Уникальный идентификатор префикса в пределах таблицы	Number
Префикс	Значение префикса русского языка, уникальное во всей таблице	String
Комментарий к префиксу	Особенности использования, написания и других характеристик префикса	String

Таблица 6.3 – Атрибуты сущности «окончание»

Атрибут	Описание	Тип данных
Номер окончания	Уникальный идентификатор окончания в пределах таблицы	Number
Префикс	Значение окончания русского языка, уникальное во всей таблице	String
Комментарий к префиксу	Особенности использования, написания и других характеристик окончания	String

Таблица 6.4 – Атрибуты сущности «соединительная гласная»

Атрибут	Описание	Тип данных
Номер	Уникальный идентификатор соединительной гласной в пределах таблицы	Number
Соединительная гласная	Значение соединительной гласной русского языка	String

Таблица 6.5 – Атрибуты сущности «суффикс»

Атрибут	Описание	Тип данных
Номер суффикс	Уникальный идентификатор суффикса в пределах таблицы	Number
Суффикс	Значение суффикса русского языка, уникальное во всей таблице	String
Комментарий к суффиксу	Особенности использования, написания и других характеристик суффикса	String

Для хранения разобранных слов необходимо ввести сущность слово (таблица 6.6).

Таблица 6.6 – Атрибуты сущности «слово»

Атрибут	Описание	Тип данных
Номер слова	Уникальный идентификатор слова в пределах таблицы	Number
Слово	Значение слова русского языка, уникальное во всей таблице	String

Для описания принадлежности слов и суффиксов той или иной части речи необходимо ввести сущность часть речи, описание атрибутов приведено в таблице 6.7.

Таблица 6.7 – Атрибуты сущности «часть речи».

Атрибут	Описание	Тип данных
Номер части речи	Уникальный идентификатор части речи в пределах таблицы	Number
Часть речи	Значение части речи русского языка, уникальное во всей таблице	String

В полученном списке существуют атрибуты, которые нельзя определить в виде одного поля БД. Такие атрибуты требуют дополнительных определений и должны рассматриваться как сущности, состоящие, в свою очередь, из атрибутов. В одном слове может содержаться несколько морфем одного типа, например, слово *перераспределение* содержит три префикса *пере-, рас-, пре-*. Для таких случаев необходимо знать номер морфемы в слове, т.е в нашем примере префиксам необходимо присвоить номера 1,2,3 соответственно. Атрибут номер морфемы в слове так же необходим для морфем корень, суффикс и соединительная гласная. Таблицы полученных сущностей приведены в таблицах 6.8-6.11.

Таблица 6.8 – Атрибуты сущности «вариант корня»

Атрибут	Описание	Тип данных
Номер корня в слове	Порядковый номер корня в слове	Number

Таблица 6.9 – Атрибуты сущности «вариант префикса»

Атрибут	Описание	Тип данных
Номер префикса в слове	Порядковый номер префикса в слове	Number

Таблица 6.10 – Атрибуты сущности «вариант суффикса»

Атрибут	Описание	Тип данных
Номер суффикса в слове	Порядковый номер суффикса в слове	Number

Таблица 6.11 – Атрибуты сущности «вариант соединительной гласной»

Атрибут	Описание	Тип данных
Номер соединительной гласной в слове	Порядковый номер соединительной гласной в слове	Number

В связи с тем, что в русском я зыке существуют слова омонимы, то одно и то же слово, одинаковое по написанию, может иметь различное значение, принадлежать разным частям речи и иметь различный морфемный разбор. Для таких ситуаций необходимо ввести еще одну сущность «слово_часть речи», атрибуты описаны в таблице 6.12.

Таблица 6.12 – Атрибуты сущности «слово_часть речи»

Атрибут	Описание	Тип данных
Номер слово_часть речи	Уникальный идентификатор части речи в пределах таблицы	Number
Комментарий	Особенности использования, написания и других характеристик слова	String
Ударение	Номер символа в слове, на которое падает ударение	Number
Оригинал	Указатель на морфологические омонимы слова	String

Для хранения вариантов разбора необходимо ввести сущность «вариант разбора», хранящая уникальный идентификатор варианта разбора (таблица 6.13).

Таблица 6.13 – Атрибуты сущности «вариант разбора»

Атрибут	Описание	Тип данных
Номер вариант разбора	Уникальный идентификатор варианта разбора в пределах таблицы	Number

Логическая схема БД представлена на рисунке 6.1.

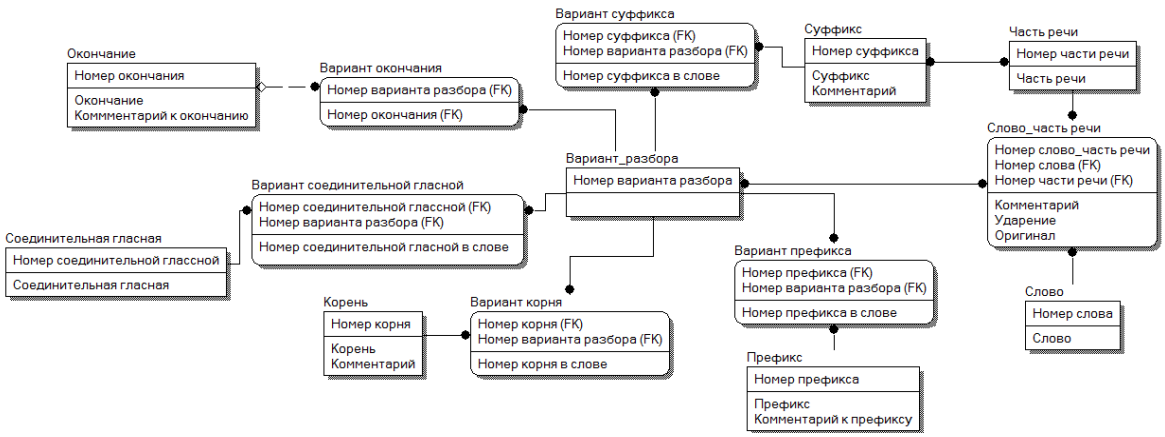


Рисунок 6.1 – Логическая модель БД

Один вариант разбора может иметь одно окончание или не иметь его вообще, но одно окончание может встречаться в нескольких вариантах разбора и обязательно встречается хотя бы в одном из разборов. Для реализации такой связи была введена сущность «вариант окончания», первичный ключ которой совпадает с первичным ключом сущности «вариант разбора», а внешний ключ с ключом сущности «окончание».

Домен можно рассматривать как подмножество значений некоторого типа данных имеющих определенный смысл. Опишем домены атрибутов сущностей. Введем сокращение Dom для определения отдельного домена. Подчеркивание будем обозначать первичные ключи.

Вариант разбора (Номер варианта разбора).

Dom (Номер варианта разбора) = {x| $A(x)$}, где $A(x)$ – истина, *если* $x = Number \ \& \ x<>x+1$, иначе – ложь.

Часть речи (Номер части речи, Часть речи)

Dom (Номер части речи) = {$x \mid B(x)$}, где $B(x)$ – истина, если $x = Number \ \& \ x<>x+1$, иначе – ложь

Dom (Часть речи) = {существительное, прилагательное, глагол, причастие, деепричастие, местоимение, числительное, наречие, союз, частица, предлог, междометие, звукоподражательные слова}

Слово (Номер слова, Слово)

Dom (Номер слова) = {$x \mid C(x)$}, где $C(x)$ – истина, если $x = Number \ \& \ x<>x+1$, иначе – ложь.

Dom (Слово) = {$x \mid D(x)$}, где $D(x)$ – истина, если $x = String \ \& \ length(x) \le 100 \ \& \ x<>x+1$, иначе – ложь,

где length(x) – длина строки.

Корень (Номер корня, Корень, Комментарий к корню)

Dom (Номер корня) = {x| $E(x)$}, где $E(x)$ – истина, если $x = Number \ \& \ x<>x+1$, иначе – ложь.

Dom (Корень) = {x| $F(x)$}, где $F(x)$ – истина, если $x = String \ \& \ length(x) \le 50 \ \& \ x<>x+1$, иначе – ложь,

где $length(x)$ – длина строки.

Dom (Комментарий к корню) = {x| $G(x)$}, где $G(x)$ – истина, если $x = String \ \& \ length(x) \le 255$, иначе – ложь,

где $length(x)$ – длина строки.

Префикс (Номер префикса, Префикс, Комментарий к префиксу)

Dom (Номер префикса) = {x| $H(x)$}, где $H(x)$ – истина, если $x = Number \ \& \ x<>x+1$, иначе – ложь.

Dom (Префикс) = {x| $I(x)$}, где $I(x)$ – истина, если $x = String \ \& \ length(x) \le 10 \ \& \ x<>x+1$, иначе – ложь,

где $length(x)$ – длина строки.

Dom (Комментарий к префиксу) = {x| $J(x)$}, где $J(x)$ – истина, если $x = String \ \& \ length(x) \le 255$, иначе – ложь,

где $length(x)$ – длина строки.

Суффикс (Номер суффикса, Суффикс, Комментарий к суффиксу)

Dom (Номер суффикса) = {x| $K(x)$}, где $K(x)$ – истина, если $x = Number \ \& \ x<>x+1$, иначе – ложь.

Dom (Суффикс) = {x| $L(x)$}, где $L(x)$ – истина, если $x = String \ \& \ length(x) \le 8 \ \& \ x<>x+1$, иначе – ложь,

где $length(x)$ – длина строки.

Dom (Комментарий к суффиксу) = {x| $M(x)$}, где $M(x)$ – истина, если $x = String \ \& \ length(x) \le 255$, иначе – ложь,

где $length(x)$ – длина строки.

Окончание (Номер окончания, Окончание, Комментарий к окончанию)

Dom (Номер окончания) = {x| $N(x)$}, где $N(x)$ – истина, если $x = Number \ \& \ x<>x+1$, иначе – ложь.

Dom (Окончание) = {x| $O(x)$}, где $O(x)$ – истина, если $x = String \ \& \ length(x) \le 5 \ \& \ x<>x+1$, иначе – ложь,

где $length(x)$ – длина строки.

Dom (Комментарий к окончанию) = {x| $P(x)$}, где $P(x)$ – истина, если $x = String \ \& \ length(x) \le 255$, иначе – ложь,

где $length(x)$ – длина строки.

Соединительная гласная (Номер соединительной гласной, Соединительная гласная)

Dom (Номер соединительной гласной) = {$1, 2$}

Dom (Соединительная гласная) = {o, e}

Слово_часть речи (Номер слово_часть речи, Комментарий, Ударение, Оригинал)

Dom (Номер слово_часть речи) = {x| $R(x)$}, где $R(x)$ – истина, если $x = Number \ \& \ x<>x+1$, иначе – ложь.

Dom (Комментарий) = {x| $S(x)$}, где $S(x)$ – истина, если $x = String \ \& \ length(x) \le 255$, иначе – ложь,

где *length(x)* – длина строки.

Dom (Ударение) = {*x*| *T(x)*}, где *T(x)* – истина, если *x = Number,* , иначе – ложь.

Dom (Оригинал) = {*x*| *W(x)*}, где *W(x)* – истина, если *x = String & length(x) ≤ 255,* иначе – ложь,

где *length(x)* – длина строки.

Вариант префикса (Номер префикса, Номер варианта разбора, Номер префикса в слове)

Домены для атрибутов Номер префикса и Номер варианта разбора описаны выше.

Dom (Номер префикса в слове) = {*x*| *U(x)*}, где *U(x)* – истина, если *x = Number*, иначе – ложь.

Вариант корня (Номер корня, Номер варианта разбора, Номер корня в слове)

Домены для атрибутов Номер корня и Номер варианта разбора описаны выше.

Dom (Номер корня в слове) = {*x*| *V(x)*}, где *V(x)* – истина, если *x = Number*, иначе – ложь.

Вариант суффикса (Номер суффикса, Номер варианта разбора, Номер суффикса в слове)

Домены для атрибутов Номер суффикса и Номер варианта разбора описаны выше.

Dom (Номер суффикса в слове) = {*x*| *Y(x)*}, где *Y(x)* – истина, если *x = Number,* иначе – ложь.

Вариант соединительной гласной (Номер соединительной гласной, Номер варианта разбора, Номер соединительной гласной в слове)

Домены для атрибутов Номер соединительной гласной и Номер варианта разбора описаны выше.

Dom (Номер соединительной гласной в слове) = {*x*| *Z(x)*}, где *Z(x)* – истина, если *x = Number*, иначе – ложь.

6.3 Физическая модель базы данных

На основании созданной логической модели создаем базу данных в СУБД MySQL.

Сущностям соответствуют таблицы БД, атрибутам сущностей – столбцы таблиц. Преобразуем домены в соответствующие типы данных, устанавливаем необходимые ограничения на значения.

В описание доменов можно выделить 4 основные функции:

1 $F_1 = \{x| A(x)\}$, где $A(x)$ – истина, *если* $x = Number \ \& \ x<>x+1$, иначе – ложь.

2 $F_2 = \{x \mid B(x)\}$, где $B(x)$ – истина, если $x = String \ \& \ length(x) \leq$ *Некоторое конкретное значение числа* $\& \ x<>x+1$, иначе – ложь

3 $F_3 = \{x| C(x)\}$, где $C(x)$ – истина, если $x = String \ \& \ length(x) \leq$ *Некоторое конкретное значение числа*, иначе – ложь,

4 $F_4 = \{x| D(x)\}$, где $D(x)$ – истина, если $x = Number$, иначе – ложь.

В СУБД MySQL этим описаниям будут соответствовать типы данных:

1 Тип Int, свойства Автоинкремен и Первичный ключ;

2 Тип Char, Длинна = Некоторое конкретное значение числа, свойство Уникальный;

3 Тип VarChar, Длинна = Некоторое конкретное значение числа, свойство Возможны пустые значения;

4 Тип Int.

После создания таблиц и описания атрибутов таблиц, создаем схему данных. На рисунке 6.2 представлена созданная схема данных в соответствии с описанными в логической модели связями. Для устранения связи многие ко многим были введены дополнительные сущности «часть речи_суффикс», «вариант разбора_часть речи». Ключи сущностей состоят из ключей сущностей «часть речи», «суффикс» и «вариант разбора», «слово_часть речи» соответственно.

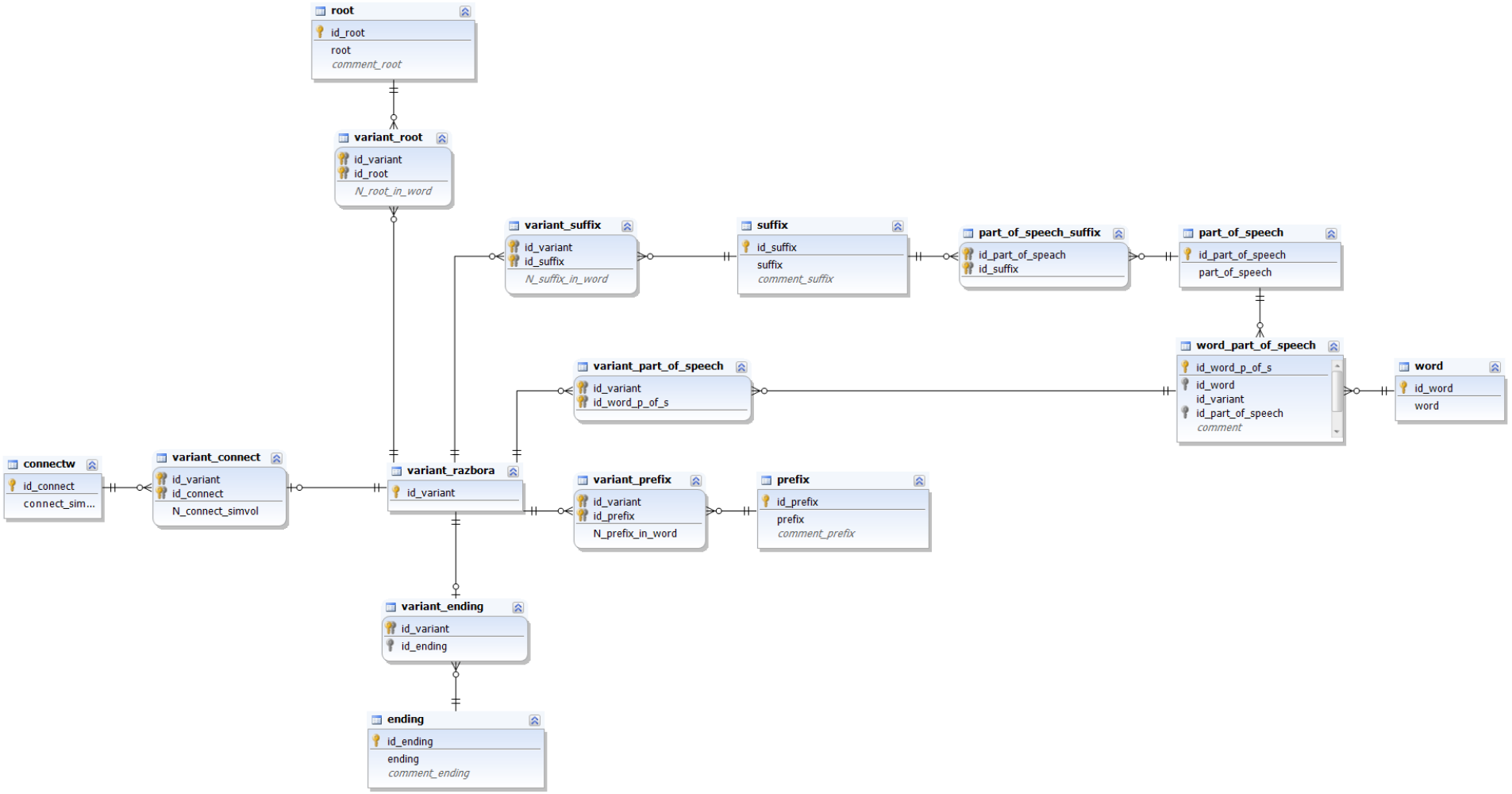


Рисунок 6.2 – Схема данных

6.4 Создание запросов

В системе морфемного анализа слов русского языка используются различные запросы к БД. Их можно разделить на три группы: запросы для непосредственного осуществления морфемного анализа, запросы для отображения данных БД при работе с ними эксперта, и запросы редактирования БД.

К первой группе запросов относятся запросы:

7 ***Запрос для поиска корней слова***

```
SELECT root
FROM
root
WHERE
'картошка' LIKE concat(root, '%')
AND (length(root) = 3
OR length(root) > 3);
```

В примере запрос находит все возможные варианты корней, которые можно выделить из начала слова *картошка*. В системе для разбора слова в запрос подставляются различные части слова, ограниченные слева, благодаря чему можно построить все варианты корней. В запросе установлено ограничение на длину корня, она должна быть больше или равна трем. Данное ограничение введено в связи с большим количеством всевозможных вариантов разбора при установлении корня меньшей длины. Результаты запроса представлены на рисунке 6.3. Тестирование запросов осуществлялось в программе dbForge Studio for MySQL [17].

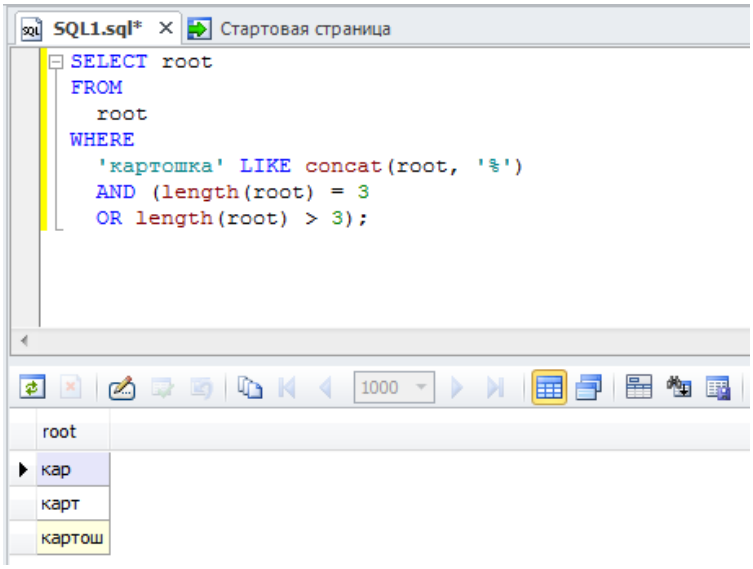


Рисунок 6.3 – Запрос для поиска корней слова

8 ***Запрос для поиска префиксов слова***

SELECT prefix

FROM

prefix

WHERE

'созвездие'

LIKE concat (prefix, '%');

В примере запрос находи все возможные варианты префиксов, с которых может начинаться слово *созвездие*. Суть такого выделения точно такая же, как и для корней. Результаты запроса представлены на рисунке 6.4.

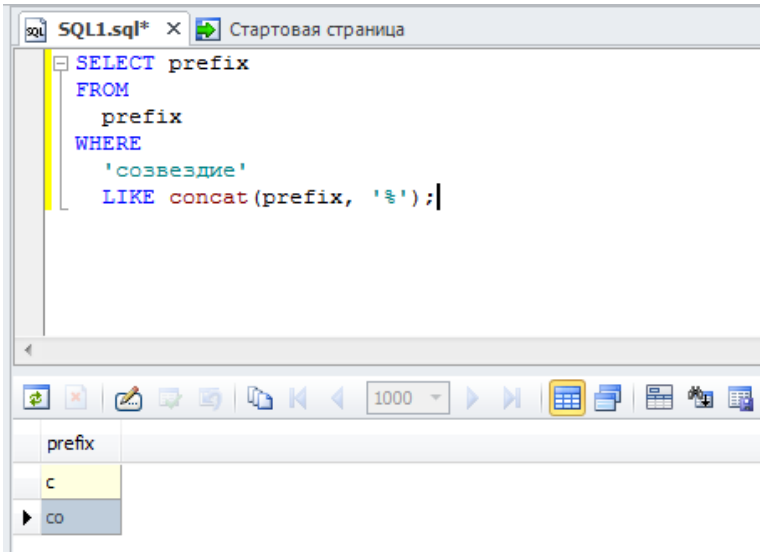


Рисунок 6.4 – Запрос для поиска префиксов слова

9 ***Запрос для поиска суффиксов слова***

SELECT suffix

FROM suffix,part_of_speech_suffix

WHERE 'енный'

LIKE Concat(suffix,'%')

AND suffix.id_suffix = part_of_speech_suffix.id_suffix

AND part_of_speech_suffix.id_part_of_speach =3;

Для суффиксов запрос более сложный, чем для префиксов и корней, так как суффиксы можно отнести к различным частям речи. Поэтому при выборке суффиксов нужно проверять относится ли он к той части речи, к которой относится разбираемое слово. Для примера рассмотрим слово *соломенный* (прилагательное, корень *солом*). Для выделения суффиксов нужно слева убрать из слова корень *солом*, после чего останется только *енный*, именно эту часть слова отправляем в запрос. В БД прилагательное имеет уникальный номер 3, поэтому необходимо выделять только те суффиксы, которые имеют в связной таблице суффиксов и

частей речи (*part_of_speech_suffix*) вторичный ключ *id_part_of_speach =3*. Результаты выполнения запроса представлены на рисунке 6.5.

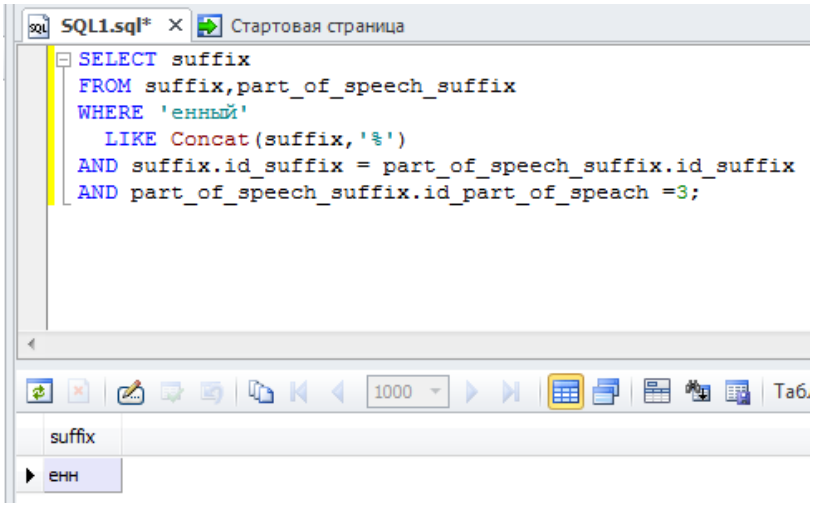


Рисунок 6.5 – Запрос для поиска суффиксов слова

10 ***Запросы для поиска вариантов разбора в БД словаря А.Н.Тихонова***

К этой же группе запросов отнесем запрос для выборки вариантов разбора слова из БД словаря А.Н.Тихнова. Запрос необходим для представления пользователю возможных вариантов разбора, которые приведены в словаре.

SELECT word

, composition

, comment

FROM

bd

WHERE

word = 'простой';

Для примера предоставим вариант разбора слова *простой* (рисунок 6.6).

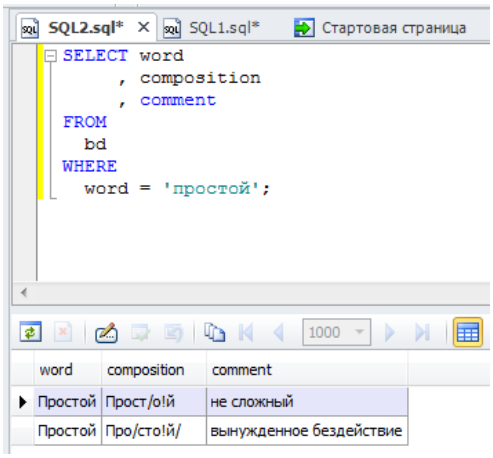


Рисунок 6.6 – Запросы для поиска вариантов разбора в БД словаря А.Н.Тихонова

Вторая группа запросов – запросы для отображения данных БД при работе с ними эксперта. К ним относятся следующие запросы:

1 ***Запрос вывода БД словаря А.Н. Тихнова***

```
SELECT *
FROM
bd;
```

В системе в данный запрос может быть добавлена одна из трех возможных частей с условием группировки *ORDER BY:*

- ORDER BY word; − сортировка по слову от А до Я;
- ORDER BY word desc; − сортировка по слову от Я до А;

2 ***Запрос выборки строк из БД А.Н. Тихонова по уникальному номеру слова***

```
SELECT word
FROM
bd
WHERE
id = 12;
```

В результате запроса будет выдано слово *аббатиса.*

3 ***Запрос вывода таблицы суффиксов***

```
SELECT *
FROM suffix;
```

4 ***Запрос вывода частей речи***

```
SELECT part_of_speech
FROM
part_of_speech
ORDER BY
id_part_of_speech;
```

Первые четыре запроса простые и не нуждается в дополнительном разъяснении.

5 ***Запрос вывода слов, имеющих информацию о разборе в БД***

```
SELECT variant_razbora.id_variant
, word
, part_of_speech
FROM
part_of_speech,
 variant_part_of_speech, variant_razbora, word, word_part_of_speech
WHERE
```

variant_razbora.id_variant = variant_part_of_speech.id_variant

AND variant_part_of_speech.id_word_p_of_s = word_part_of_speech.id_word_p_of_s

AND word.id_word = word_part_of_speech.id_word

AND word_part_of_speech.id_part_of_speech = part_of_speech.id_part_of_speech;

Запрос выводит информацию о разобранных словах из связных таблиц *part_of_speech, variant_part_of_speech, variant_razbora, word, word_part_of_speech* (рисунок 6.7). В системе в данный запрос может быть добавлена одна из трех возможных частей с условием группировки *ORDER BY:*

- ORDER BY word; − сортировка по слову от А до Я;
- ORDER BY word desc; − сортировка по слову от Я до А;
- ORDER BY part_of_speech; − сортировка по части речи.

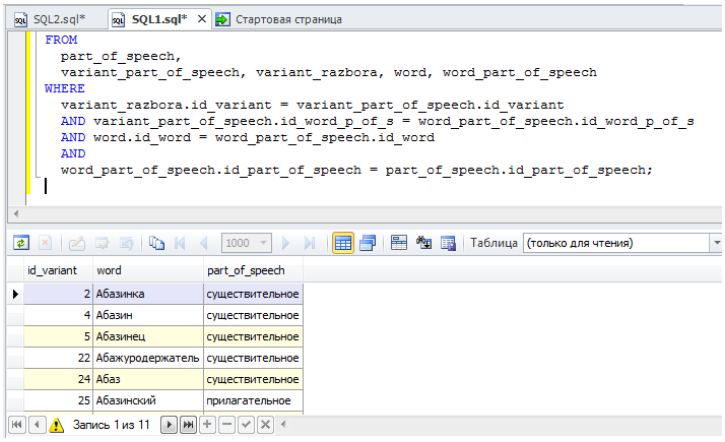


Рисунок 6.7 – Запрос вывода слов, имеющих информацию о разборе в БД

6 ***Запрос поиска слова в таблице разобранных слов***

SELECT variant_razbora.id_variant,word,part_of_speech

FROM part_of_speech,

variant_part_of_speech, variant_razbora, word, word_part_of_speech

WHERE variant_razbora.id_variant = variant_part_of_speech.id_variant

AND variant_part_of_speech.id_word_p_of_s = word_part_of_speech.id_word_p_of_s

AND word.id_word = word_part_of_speech.id_word

AND word_part_of_speech.id_part_of_speech = part_of_speech.id_part_of_speech

AND word = 'абазин';

Запрос отличается от предыдущего лишь добавлением условия для поиска конкретного слова в таблице. Примере в таблице ищется слово *абазин* (*word = 'абазин'*).

7 ***Запрос на выборку слова из таблицы разобранных слов по уникальному номеру его разбора***

SELECT word,part_of_speech

FROM word, word_part_of_speech, variant_razbora,

variant_part_of_speech,part_of_speech

WHERE word.id_word = word_part_of_speech.id_word

AND word_part_of_speech.id_word_p_of_s=variant_part_of_speech.id_word_p_of_s

AND variant_part_of_speech.id_variant=variant_razbora.id_variant

AND word_part_of_speech.id_part_of_speech=part_of_speech.id_part_of_speech

AND variant_razbora.id_variant =38;

В системе в запрос номер варианта разбора подставляется в зависимости от номера строки таблицы слов, имеющих разбор. Для примера введен номер 38, которому в таблице соответствует слова *абака* (рисунок 6.8).

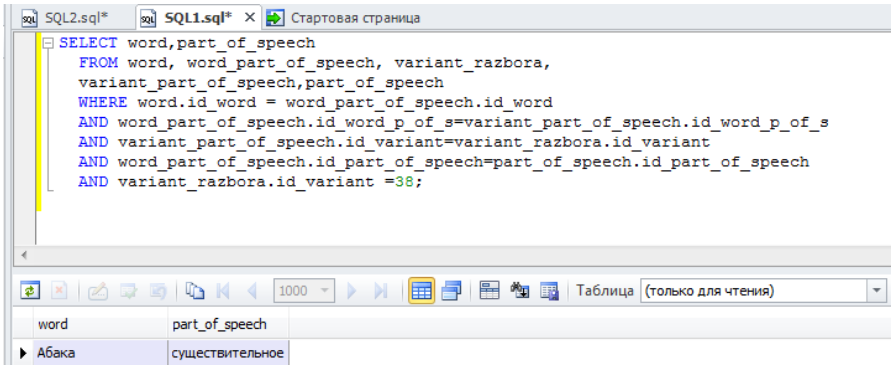


Рисунок 6.8 – Запрос поиска слова по его уникальному номеру разбора

8 ***Запрос выборки уникального номера частей речи, к которым относится суффикс***

SELECT id_part_of_speach

FROM

part_of_speech_suffix

WHERE

id_suffix = 4;

Результаты запроса представлены на рисунке 6.9. Запрос необходим для визуализации отношений суффиксов и частей речи. Уникальному номеру суффиксов 4 соответствует суффикс *-ан-*, который в свою очередь встречается у таких частей речи как существительное (*id_part_of_speach=1*) и прилагательное (*id_part_of_speach=3*).

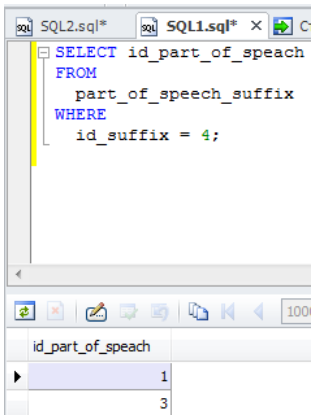


Рисунок 6.9 – Запрос выборки уникального номера частей речи, к которым относится суффикс

Последняя группа запросов – запросы редактирования БД:

1 ***Запрос вставки нового варианта разбора***

INSERT INTO variant_razbora VALUES (null);

Для сохранения нового варианта разбора для начала сохраняется его уникальный номер в таблице *variant_razbora.* Данное поле является первичным ключом таблицы и задано как автоинкремент, поэтому записывается значение *null*, а СУБД сама вписывает необходимое значение автоинкремента в таблицу.

2 ***Запрос получения значения автоинкрементна***

SELECT @@IDENTITY;

Запрос необходим для сохранения значения варианта разбора в связных таблицах.

3 ***Запрос выборки уникального номера части речи***

SELECT id_part_of_speech

FROM

part_of_speech

WHERE

part_of_speech = ‘существительное’;

Запрос прост и не требует дополнительного описания. Запрос необходим для сохранения значения части речи нового слова в связных таблицах *part_of_speech_suffix, word_part_of_speech.*

4 ***Запрос выборки уникального номера слова***

SELECT id_word

FROM

word

WHERE

word = ‘слово’;

5 ***Запрос вставки значения в таблицу word_part_of_speech***

INSERT INTO word_part_of_speech (id_word, id_part_of_speech) VALUES (1,1);

Конкретное значение *id_word* определяется либо вызовом запроса 2, этой группы запросов, либо вызовом запроса 4. Конкретное значение *id_part_of_speech* определяется вызовом запроса 3. В примере вписаны значений 1,1.

6 ***Запрос вставки данных в таблицу variant_part_of_speech***

INSERT INTO variant_part_of_speech VALUES (1, LAST_INSERT_ID());

Конкретное значение *id_ variant* определяется вызовом запроса 2. Уникальный номер таблицы *word_part_of_speech* определяется функцией *LAST_INSERT_ID*, возвращающей идентификатор последней вставки.

7 ***Запрос вставки данных в таблицу word***

INSERT INTO word (word) VALUES (‘слово’);

Запрос выполняется при отсутствии слова в БД.

8 ***Запросы сохранения префикса при добавлении нового слова в БД***

Сперва выполняется запрос выборки для определения идентификатора приставки, выделенной в слове, предположим это приставка *пре-*:

SELECT id_prefix

FROM

prefix

WHERE

prefix = ‘пре’;

Далее выполняется непосредственная вставка данных в таблицу:

INSERT INTO variant_prefix VALUES (1, 1, 1);

Первое значение это идентификатор разбора, второе уникальный номер префикса, третье – номер префикса в слове.

При отсутствии приставки в таблице, она добавляется в БД.

INSERT INTO prefix (prefix) VALUES (‘пре’);

9 ***Запросы сохранения корня при добавлении нового слова в БД***

Для корня запросы охранения действуют точно так же, как для префикса.

SELECT id_root

FROM

root

WHERE

root = ‘кот’;

Далее выполняется непосредственная вставка данных в таблицу:

INSERT INTO variant_root VALUES (1, 1, 1);

Первое значение это идентификатор разбора, второе уникальный номер корня, третье – номер корня в слове.

При отсутствии корня в таблице, она добавляется в БД.

INSERT INTO root (root) VALUES (‘ кот’);

10 ***Запросы сохранения соединительной гласной при добавлении нового слова в БД***

Работает аналогично запросам с корнем и префиксом. Но не может возникнуть запроса, при котором соединительную гласную необходимо добавлять в БД, так как в русском языке всего две соединительных гласных: *-о-*, *-е-*.

SELECT id_connect

FROM

connectw

WHERE

connect_simvol = 'o';

Далее выполняется непосредственная вставка данных в таблицу:

INSERT INTO variant_connect VALUES (1,1,1);

11 ***Запросы сохранения суффикса при добавлении нового слова в БД***

При сохранении суффикса используются запросы аналогичные запросам при добавлении корня и префикса.

SELECT id_suffix

FROM

suffix

WHERE

suffix = 'ик';

Далее выполняется непосредственная вставка данных в таблицу:

INSERT INTO variant_ suffix VALUES (1,1,1);

При отсутствии связи между суффиксом и частью речи необходимо добавить значение в соответствующую таблицу.

INSERT INTO part_of_speech_suffix VALUES (1,1);

Значения 1,1 – это уникальные номера части речи и суффикса соответственно.

При отсутствии корня в таблице, она добавляется в БД.

INSERT INTO suffix (suffix) VALUES (‘ик’);

12 ***Запрос сохранения окончания при добавлении нового слова в БД***

SELECT id_ending

FROM

ending

WHERE

ending = 'ся';

INSERT INTO variant_ending VALUES (1,1).

Окончания в слове либо нет совеем, либо оно одно, поэтому поля номер окончания в слове в таблице нет.

INSERT INTO ending (ending) VALUES (‘ся’);

13 ***Запросы удаления варианта разбора слова***

DELETE FROM variant_root WHERE id_variant=1;

DELETE FROM variant_prefix WHERE id_variant=1;

DELETE FROM variant_connect WHERE id_variant=1;

DELETE FROM variant_suffix WHERE id_variant=1;

DELETE FROM variant_ending WHERE id_variant=1;

DELETE FROM variant_part_of_speech WHERE id_variant=1;

DELETE FROM variant_razbora WHERE id_variant=1;

Для удаления одного варианта разбора необходимо последовательно удалить информацию из всех таблиц, в которых он хранится, т.е из таблиц *variant_root, variant_prefix, variant_connect, variant_suffix, variant_suffix, variant_ending, variant_part_of_speech, variant_razbora.*

Для редактирования, вставки и удаления таблиц морфем используются те же принципы. Поэтому нет смысла подробно на них останавливаться.

7 Тестирование системы

Рассмотрим, сколько вариантов разбора может предложить разработанная система для слов *Созвездие* и *Разделительный*. Установим правило, что минимальная длинна корня равна трем символам. Морфемный-орфографический словарь А. Н. Тихонова для слова *созвездие* дает следующий разбор: *Со/звезд/и/е* [4]. Знаком «/» в данном случае разделены между собой морфемы. Можем предположить, что *со* – приставка, *звезд* – корень, *и* – суффикс, *е* – окончание. Морфологический анализ слова и деревья с возможными вариантами разбора, полученные системой для слова *созвездие*, приведены на рисунке 7.1-7.2.

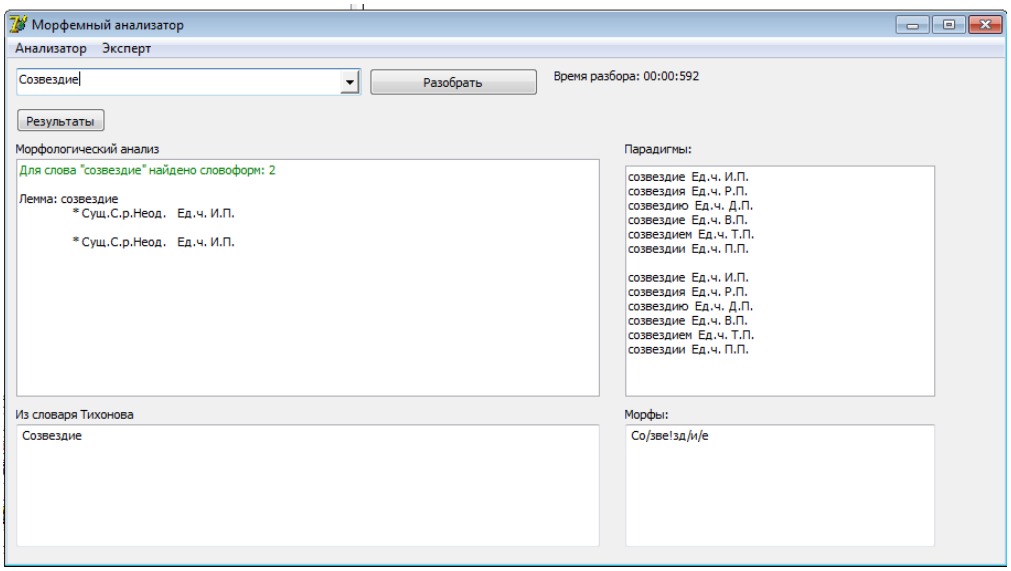


Рисунок 7.1 – Морфологический разбор слова «Созвездие»

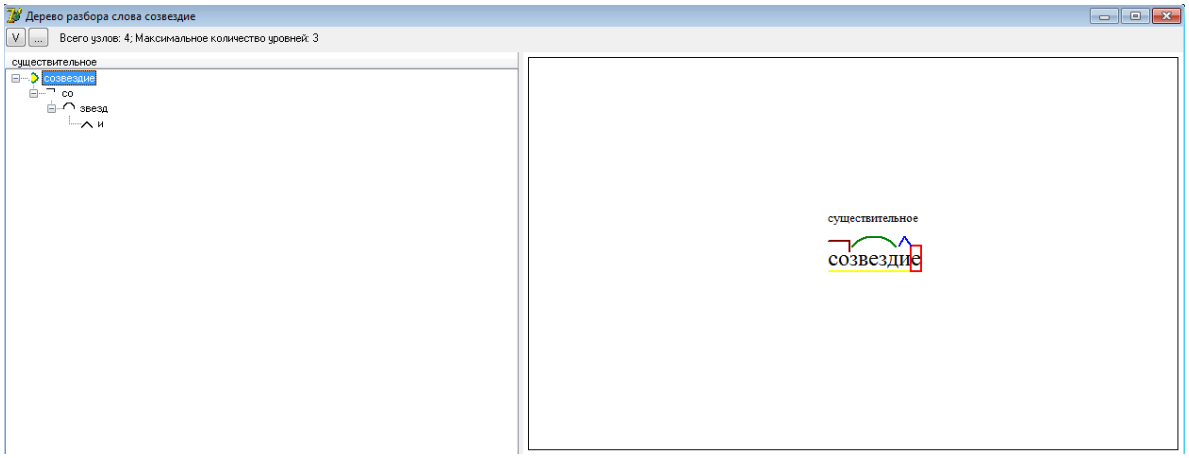


Рисунок 7.2 – Деревья разбора слова «Созвездие»

Слово «Разделительный» в словаре А.Н. Тихонова имеет следующий разбор: *Раз/дел/и/тельн/ый* (*раз* – приставка, *дел* – корень, *и* – суффикс, *тельн* – суффикс, *ый* – окончание). Разбор системы морфемного анализа представлен на рисунке 7.3 – 7.4.

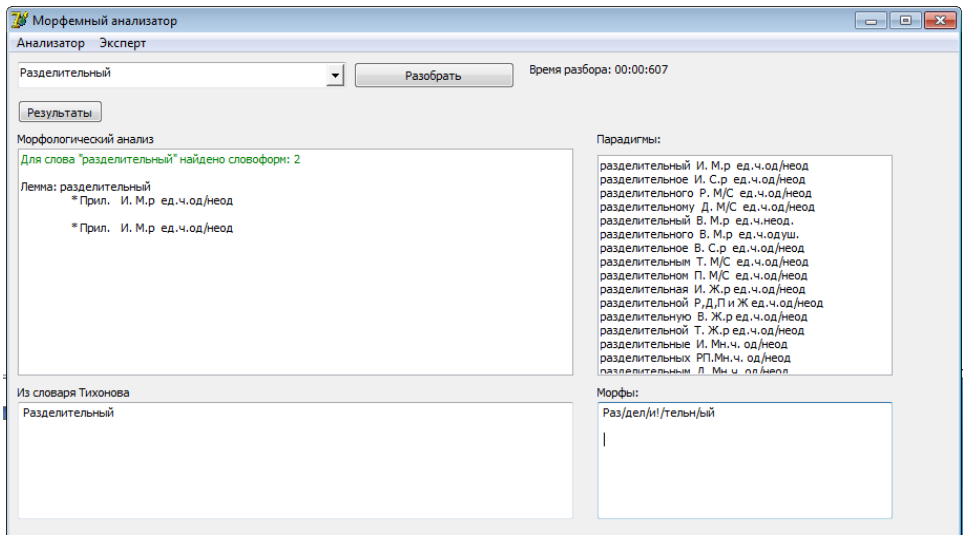


Рисунок 7.3 – Морфологический разбор слова «Разделительный»

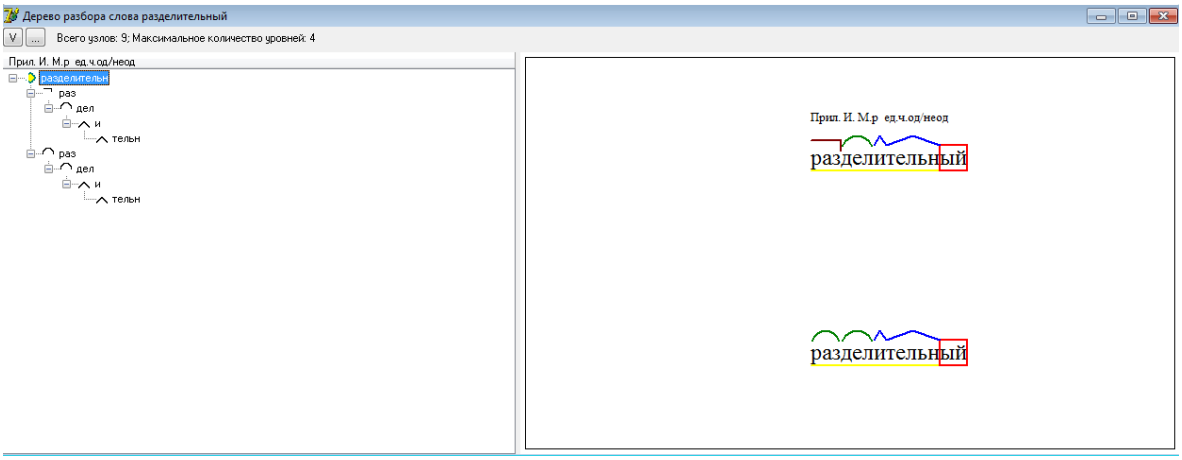


Рисунок 7.4 – Деревья разбора слова «Разделительный»

Эксперт системы может редактировать базу данных системы: добавлять морфемы, устанавливать новые связи между суффиксами и частью речи. Для установления данной связи нужно добавить или снять указатель возле части речи под таблицей суффиксов. На рисунке суффиксу *-а-* соответствуют части речи: глагол, наречие, деепричастие (рисунок 7.5).

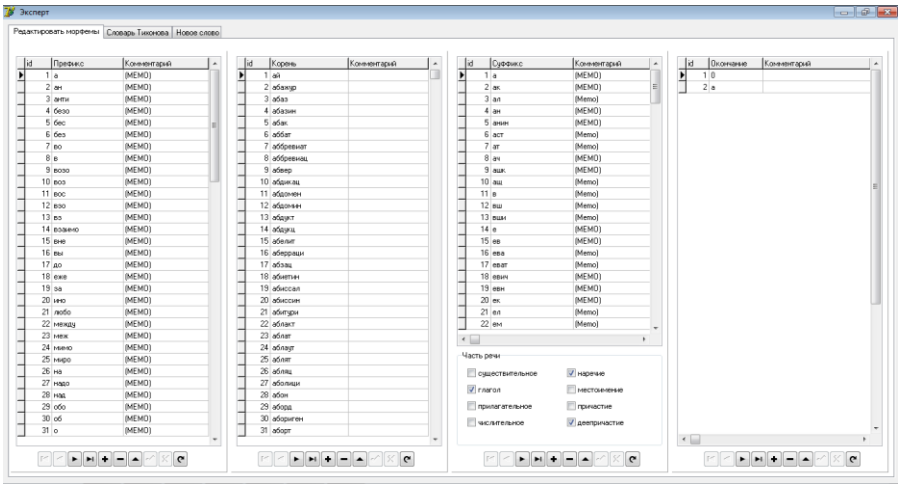


Рисунок 7.5 – Форма редактирования морфем

Эксперт может добавлять новые слова в систему на основании морфемно-орфографического словаря А.Н.Тихонова. Добавим в систему слово *Абажур* (рисунок 7.6). Разбор слова *Абажур* в словаре дано как *Абажу!р/*, то есть слово полностью состоит из корня и нулевого окончания, знаком *!* обозначено ударение на букву *У*. Таким образом, выберем в выпадающем списке напротив поля Корень значение *Абажур*, в поле окончание необходимо вписать значение 0, в выпадающем списке выбрать часть речи – существительное.

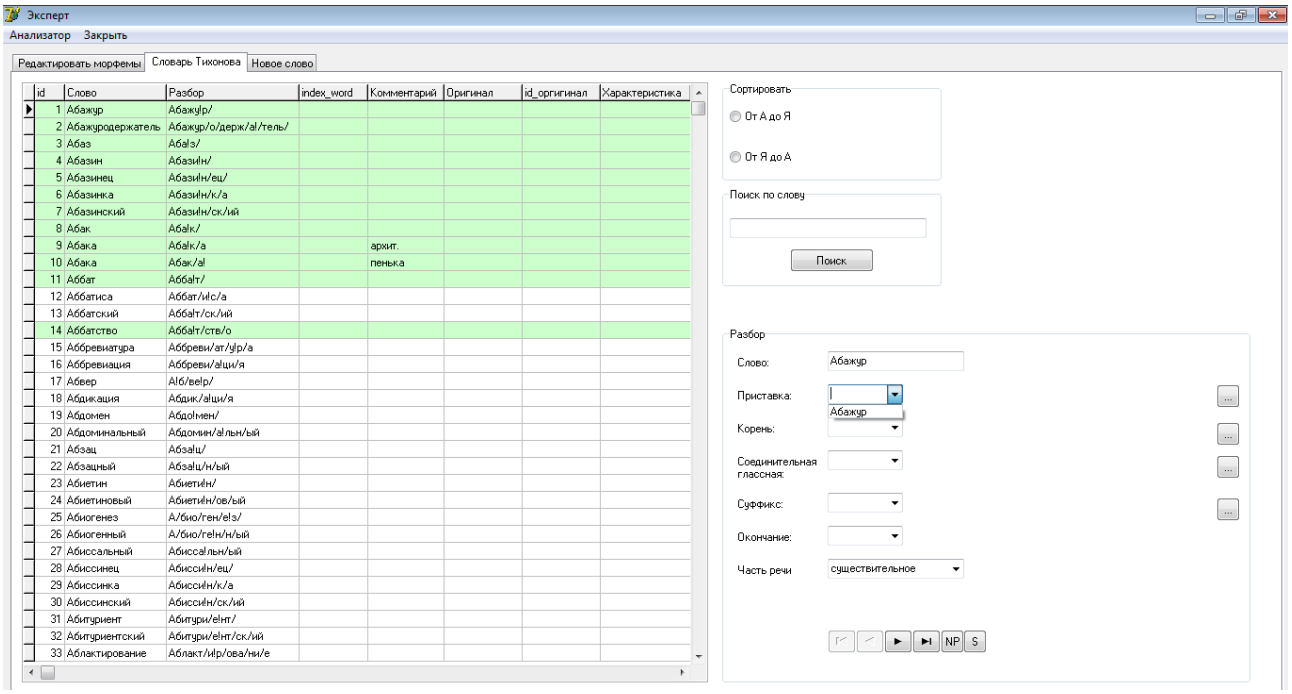


Рисунок 7.6 – Добавление слова *Абажур*

Для сохранения в БД необходимо нажать кнопку S. Кнопка NP очищает поля ввода. Слова, для которых уже сохранен разбор в системе, в таблице Словарь А.Н.Тихонова выделены зеленым цветом.

Если открыть вкладку новое поле, то видно введенное нами слово (рисунок 7.7).

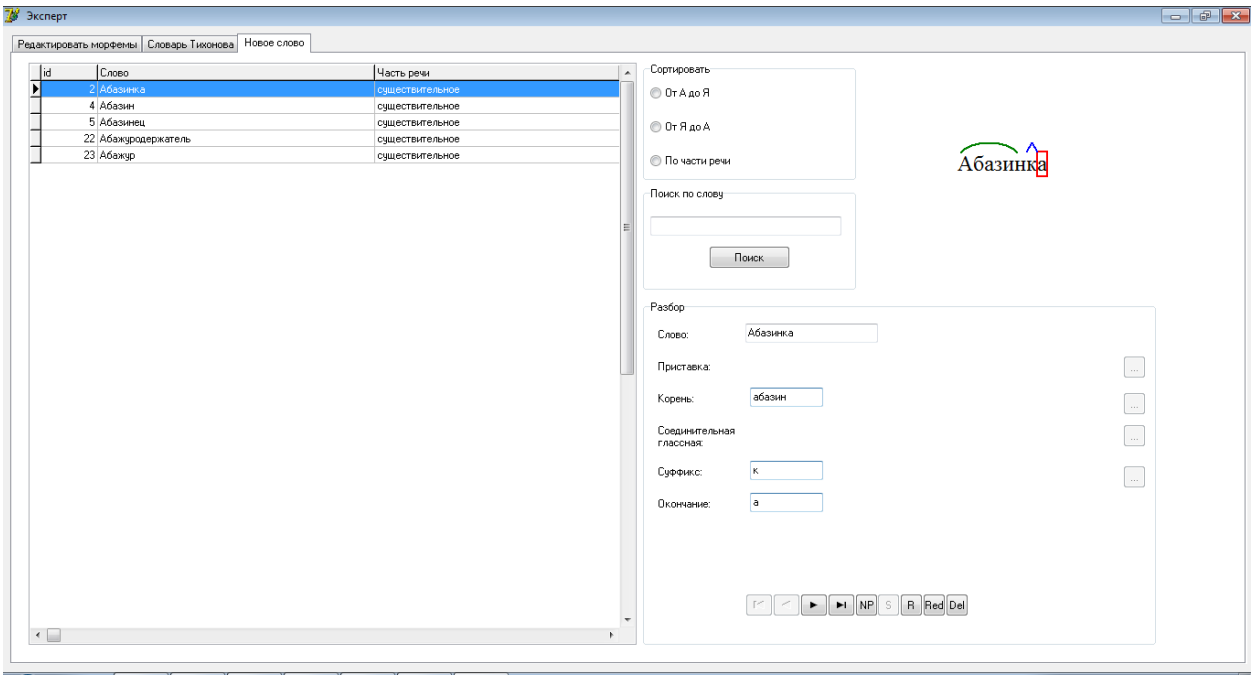


Рисунок 7.7 – Вкладка «Новое слово»

На данной вкладке, перемещаясь по списку слов, можно видеть их разбор в полях ввода. Для редактирования информации необходимо нажать кнопку Red. Для удаления слова – кнопку Del. Для добавления нового не словарного слова (без предоставления варианта разбора по словарю А.Н.Тихонова) необходимо нажать кнопку NP. При нажатии возле каждой морфемы появится поле для ввода и станут доступны кнопки S (сохранить) и кнопки добавления полей для ввода морфем.

Значения таблиц разобранных слов (рисунок 7.6) и словаря А.Н.Тихонова (рисунок 7.7) можно сортировать и осуществлять поиск по слову. Если слово не найдено, то для возвращения всех значений необходимо очистить поле поиска и еще раз нажать кнопку Поиск.

8 Проблемы и задачи, выявленные в ходе разработки системы

При изучении данной темы был выявлен ряд проблем. Так, при данном алгоритме разбора может быть получено множество вариантов разбора одного и того же слова. В русском языке нет строгих правил о том, какое минимальное или максимальное количество морфем допустимо в слове. К примеру, в слове *Созвездие* компьютер выделил как одну приставку *со-* так и сочетание приставок *с-* и *о-*. Что приводит к неопределенности. Такая многозначность встречается и с другими видами морфем.

Кроме того, компьютер не может выбрать какая из морфем приоритетнее. То есть, если морфемы совпадают по написанию, как в примере со словом *Разделительный* (*раз* может быть как корнем, так и приставкой), компьютер предлагает два варианта разбора, что тоже ведет к неоднозначности.

Русский язык – живой язык, он изменяется и развивается каждый день. Поэтому словари морфем русского языка требуют доработки и обновления. Особенно остро стоит вопрос формирования словаря корней. В русском языке огромное количество всевозможных корней, которые необходимо вносить в базу данных корней для более стабильной и правильной работы морфемного анализатора.

Так же проблемой является различная трактовка учеными некоторых морфем. К примеру, в глаголах в форме инфинитива по одним источникам *ть-* на конце слова это суффикс (В.В. Бабайцева, Д.Э.Розенталь) [13], по другим – это окончание (С.Г. Бархударова, С.Е. Крючкова, Л.Ю.Максимова, Л.А. Чешко) [13]. Кроме того стоит вопрос каким образом выделять при разборе сложных слов соединительные гласные *-о-* и *-е-* и другие подобные вопросы.

При решении вопроса о разборе слов с дефисом была обнаружена следующая проблема: при разборе сложных слов, таких как *ЖЕНЩИНА-КОШКА, УГОЛЬНО-ЧЕРНЫЙ* и т.п то, во-первых, приходится строить разбор двух независимых слов, во-вторых, если первое слово будет иметь N вариантов разбора, а второе слово после дефиса M вариантов разбора, то всего всех возможных сочетаний будет N×M. При значении N и M равным трем это 9 вариантов разбора. Это достаточно много и будет больше путать пользователя, чем поможет определить правильный вариант.

Слова с выпадающей гласной и чередующейся согласной система обрабатывать не сможет, их нужно в первую очередь внести в БД. Так как анализировать в каких случаях буква выпадает (чередуется), а в каких нет, не представляется возможным. Так как в русском языке для этого нет строгих правил, для которых можно было бы построить математические модели.

Не достаточно информации в том случае, когда необходимо выделять нулевое окончание. Нулевое окончание выделяется [8]:

- у существительных в форме именительного падежа, единственного числа, мужского рода (2 склонения) и женского рода (3 склонения), но Зализняк не позволяет определить склонение);
- у части существительных в форме родительного падежа, множественного числа, однако говорится только о части существительных, то есть ко всем словам удовлетворяющим параметрам, нельзя применить это правило;
- у кратких прилагательных и причастий в форме единственного числа, мужского рода (для кратких прилагательных в Зализняке не дается значение числа);
- у глаголов в форме прошедшего времени, единственного числа, мужского рода;
- у притяжательных прилагательных с суффиксом *-ий-* (притяжательные прилагательные так же не выделяются в Зализняке);
- у глаголов в повелительном наклонении, где нулевым окончанием выражается значение единственного числа.

Таким образом, выделить нулевое окончание можно только в 4-ом и 6-ом случае.

На данный момент система ограничивает длину корня тремя символами, так как при отсутствии данного ограничения варианты разбора слова резко возрастают, уменьшить их количество нельзя в связи с проблемой отсутствия информации о максимальном количестве морфем в слове, описанной выше. Таким образом, слова имеющим корень из одной буквы (выйти, шёл, шла, шли, шло, кого, чего и т.д.) или из двух букв системой разобраны не будут или будут разобраны неверно.

В ходе изучения темы, была рассмотрена монография Л.Г.Зубковой «Принцип знака в системе языка» [23]. Одна из глав монографии посвящена морфемному строению слова в семитологическом аспекте. В целом монография посвящена выявлению связи между звучанием и значением во внутреннем строе языка. Тема отлична от поставленной в данной дипломной работе, но в работе Л.Г Зубковой морфема рассматривается как одна из важнейших характеристик, раскрывающих принцип знака. В связи, с чем большое внимание уделено выявлению типовых морфемных структур (корневые, суффиксальные, префиксальные, префиксально-суффиксальные и т.п.) и их конкретных реализаций в виде отдельных морфемных моделей. И именно эта информация интересна и имеет практическую пользу для изучаемой темы автоматизации морфемного анализа.

В работе Л.Г Зубковой анализируется морфемное строение слов различных языков, помимо этого автор пытается выявить характер связи между морфемной структурой слова и типом текста и определить текстообразующий и тексторазличительный потенциал различ-

ных классов слов. Рассмотрим данные полученные для русского языка. Для понимания дальнейшей информации необходимо ввести ряд сокращений, используемых в монографии:

УХТ – устные художественные тексты;

ПХТ – письменные художественные тексты;

НТ – научные тексты;

РТ – разговорные тексты.

Авторам рассматривается закономерность распределения морфемных структур различной сложности в зависимости от типа языка (арабский, английский, китайский, йоруба и т.д). Для дипломной работы наиболее интересна информация о количестве морфем в слове и степени распространенности морфемной сложности для русского языка в УХТ и ПХТ (таблица 8.1-8.2).

Таблица 8.1 − Распределение классов слов по степени морфемной сложности в УХТ (без Ø)(в % от общего числа слов в данном классе)

Класс слов	Русский						
	кол-во морфем в слове						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Служебные	**98,7**		1,3				
Местоимения	41,7	**50,9**	7,4				
Собственно-знаменательные	11,4	**37,8**	**27,8**	17,2	4,8	0,9	0,1
Существительные	18,1	**60,0**	15,9	4,4	1,1	0,5	
Глаголы	0,6	16,9	**36,9**	**33,9**	9,7	1,8	0,2
Знаменательные	14,7	**39,2**	25,6	15,3	4,3	0,8	0,1
«Слово вообще»	**36,4**	29,2	19,3	11,3	3,2	0,6	0,07

Таблица 8.2 − Распределение классов слов по степени морфемной сложности в ПХТ (с Ø) (в % от общего числа слов в данном классе)

Класс слов	Русский					
	кол-во морфем в слове					
	1	**2**	**3**	**4**	**5**	**6**
Служебные	**99,3**		0,7			
Местоимения	7,6	**88,6**	3,8			
Собственно-знаменательные	3,0	**42,4**	29,0	13,8	9,4	2,4
Существительные		**67,5**	28,3		4,2	
Глаголы		10,1	**34,3**	**30,3**	21,2	4,1
Знаменательные	4,0	**52,1**	23,7	11,0	7,4	1,8
«Слово вообще»	29,1	**38,5**	17,6	8,0	5,4	1,4

На основе этих данных можно говорить, что в русском языке сложная морфемная структура слов. Наиболее распространена морфемная структура, состоящая из двух, трех, в меньшей степени, четырех морфем в слове. Используя данную информацию возможен анализ деревьев разбора, полученных разработанной системой морфемного анализа. И, с проведением математических расчетов вероятности наиболее часто встречающихся морфемных структур, возможно с достаточно большей точностью исключать из деревьев разбора неверные и наименее вероятные варианты разбора слова.

Кроме этого, монографии Л.Г.Зубкова рассматривает вопрос типовых морфемных структур и конкретных морфемных моделей имен существительных и глаголов. Ею рассматриваются только две части речи, потому что ведется анализ между различными языками, а глагол и существительное являются так сказать «универсальными» частями речи в том или ином виде существующих во всех языках.

В дипломной работе так же будет рассмотрена только информация, полученная для русского языка. Зубковой введены следующие сокращения для описания типовых морфемных структур: К — корень, П — префикс, С — суффикс, Со — суффикс словообразовательный, Сф — суффикс формообразующий, Си — суффикс словоизменительный, Ф — флексия, ин — интерфикс, Пф — постфикс, Ø — нулевая морфема.

При рассмотрении морфемных моделей автор приходит к следующим выводам: «если ограничиться анализом морфемных моделей с частотой не менее 1% от числа словоформ данной части речи в данном тексте, то у существительных таких моделей всего 10, а у глаголов — 13. Если же исключить также модели, встретившиеся только в одном или двух текстах, то число типичных моделей еще более сократится. (Исключению подлежали у существительных модели ПКССФ и КСССФ в НТ, КинКФ и ККСФ в ПХТ, ПКСССФ в УХТ; у глаголов — модели КСССФ в УХТ и ПХТ, КФПф в НТ и РТ, ПКФПф в РТ.) Таким образом, у существительных остается 5 моделей, у глаголов — 10» [23]. Иерархия типичных морфемных моделей для существительных и глаголов приведена в таблицах 8.3, 8.4 соответственно.

Таблица 8.3 − Иерархия типичных морфемных моделей существительных в разных текстах (в % от общего числа существительных в данном тексте)

УХТ			РТ			ПХТ			НТ		
1.	КФ	73,2	1.	КФ	64,7	1.	КФ	61,2	1.	КФ	56,6
2.	КСФ	14,6	2.	КСФ	20,3	2.	КСФ	21,3	2.	КСФ	22,5
3.	КССФ	8,7	3.	КССФ	3,9	3.	ПКФ	4,5	3.	ПКСФ	7,9
4.	ПКФ	1,5	4.	ПКФ	3,0	4.5.	КССФ	3,0	4.	КССФ	4,2
5.	*ПКСССФ*	1,1	5.	ПКСФ	1,9	4.5.	ПКСФ	3,0	5.	*ПКСССФ*	2,9
6.5.	ПКСФ	0,55				6.	*КинКФ*	2,3	6.	ПКФ	1,4
6.5.	*ПКССФ*	0,55				7.	*ККСФ*	1,6	7.	*КСССФ*	1,2

Таблица 8.4 − Иерархия морфемных моделей глаголов в разных текстах(в % от общего числа глаголов в данном тексте)

УХТ			ПХТ			НТ			РТ		
1.	ПКСФ	22,3	1.	ПКССФ	20,4	1.	КФ	22,3	1.	КФ	25,4
2.	ПКССФ	16,4	2.	КССФ	17,9	2.	КСФ	19,6	2.	КСФ	22,7
3.	КФ	14,8	3.	КСФ	13,8	3.	ПКСФ	18,6	3.	ПКСФ	11,7
4.	КСФ	13,9	4.	ПКСФ	13,0	4.	ПКСФПф	9,5	4.	КССФ	10,4
5.	КССФ	10,0	5.	ПКССФПф	8,3	5.	КСФПф	8,1	5.	ПКФ	9,8

Продолжение таблицы 8.4

УХТ			ПХТ			НТ			РТ		
6.	ПКФ	8,9	6.	КФ	7,5	6.	ПКФ	3,9	6.	ПКССФ	7,4
7.	ПКССФПф	3,9	7.	КССФПф	6,1	7.	КССФ	3,6	7.	ПКСФПф	2,7
8.	ПКСФПф	2,0	8.	ПКФ	5,1	8.	*КФПф*	2,5	8.	ПКССФПф	2,1
10.	КСФПф	1,1	9.	ПКСФПф	1,7	9.	КССФПф	2,3	9.	КСФПф	1,9
10.	КССФПф	1,	10.	*КССССФ*	1,3	10.	ПКССФПф	1,9	10.5.	КССФПф	1,7
10.	*КССССФ*	1,1	11.	КСФПф	1,2	11.	ПКССФ	1,7	10.5.	*КФПф*	1,7
									12.	*ПКФПф*	1,1

Для существительных, во всех исследуемых текстах, наиболее распространены модели КФ и КСФ (первый и второй ранги). Со снижением частоты встречаемости слова в тексте сложность морфемной модели существительных увеличивается. В глаголах степень сложности моделей, занимающих самые высокие ранги — первый, второй, третий, в нехудожественных текстах со снижением частоты возрастает, а в художественных, прежде всего в письменных, убывает.

На основании этих данных в разработанную систему можно ввести функцию расчета наиболее вероятных морфемных моделей слов для существительных и глаголов. Что поможет пользователю наиболее верно выбирать вариант разбора при отсутствии его в системе и предоставлении системой нескольких возможных вариантов разбора. Такая возможность повысит качество программного продукта, за счет более точного анализа слова и предоставления пользователю отсортированных вариантов разбора по степени наиболее частой встречаемости морфемных моделей. С большой степенью точности, можно будет определить какой из нескольких вариантов разбора, полученных системой, верный.

В связи с тем, что информация, полученная из монографии Л.Г.Зубковой, была найдена и рассмотрена на последних этапах написания дипломного проекта, функции, которые описаны выше, не введены в систему. Поэтому при дальнейшей разработке системы, в первую очередь, необходимо определить с помощью каких математических средств, возможен расчет вероятности наиболее верных результатов разбора. Разработка и добавление выше описанных функций, одна из важнейших целей при дальнейшей разработке системы.

Таким образом, предполагается, что в последующих работах будут найдены решения проблем, описанных в данной главе. Разработанная система будет доработана и повышено качество результатов морфемного разбора системы.

ЗАКЛЮЧЕНИЕ

В результате выполнения дипломной работы был разработан и создан программный продукт, автоматизирующий морфемный анализ слов русского языка.

В первой главе пояснительной записки проведен обзор методов и систем морфемного анализа русскоязычных текстов. Были изучены два основных подхода к разбору слова по составу: формально-структурный и формально-смысловой. Систем разбора слов по составу найдено не было.

Были изучены правила морфемного анализа, собран словарь морфем. Разработан алгоритм морфемного анализа слов русского языка.

Для определения части речи, выделения основы слова и окончания были изучены и внедрены в систему библиотека MCR.dll и программа mysem от Яндекс.

Особое внимание было уделено таким модулям системы, как: сокращение вариантов разбора слов, визуализация результатов разбора, экспертное редактирование БД системы.

БД системы содержит информацию о морфемах и их признаках, а так же морфемные словари и результаты разбора слов. Проектированию БД уделены главы 6.2-6.4 пояснительной записки.

Созданная система протестирована, устранены явные ошибки.

Разработано руководство пользователя и эксперта

Разработка системы остается актуальной и перспективной задачей. Необходимо провести дополнительный анализ выявленных в ходе разработки проблем (пункт 8 пояснительной записки) и внести необходимые изменения в систему; поработать над интерфейсом программы.

Таким образом, достигнуты поставленные цели и сформированы новые задачи для дальнейшей работы.

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1. Филиппова Л.С. Современный русский язык. Морфемика. Словообразование: учеб. пособие. / Л.С. Филиппова. — М.: Флинта: Наука, 2009. — 248 с. (Дата обращения: 11.09.12)
2. И.П.Суслов Введение в теоретическое языкознание. Модуль 4. Основы общей морфологии. Принципы морфемного анализа. [Электронный ресурс] // ИПС [Сайт] URL: http://homepages.tversu.ru/~ips/4_02.htm (Дата обращения: 12.09.12)
3. Елена Литневская Русский язык: краткий теоретический курс [Электронный ресурс] // Электронная библиотека [Сайт] URL: http://bookz.ru/authors/elena-litnevskaa/russkii-_472/page-5-russkii-_472.html (Дата обращения: 12.09.12)
4. Тихонов А. Н. Морфемно-орфографический словарь [Электронный ресурс] // Яндекс Словари [Сайт] URL: http://slovari.yandex.ru/~книги/Морфемно-орфографический%20словарь/ (Дата обращения: 12.09.12)
5. Живой Корнеслов [Электронный ресурс] // oomnik лингвотехнология [Сайт] URL: http://www.oomnik.ru/korneslov/ (Дата обращения: 13.09.12)
6. С.А.Старостина Морфологический анализатор [Электронный ресурс] // [Сайт] URL: http://starling.rinet.ru/morph.htm (Дата обращения: 13.09.12)
7. О программе mystem [Электронный ресурс] // Яндекс компания [Сайт] URL: http://company.yandex.ru/technologies/mystem/ (Дата обращения: 14.03.2013)
8. Бабайцева В.В., Чеснокова Л.Д. Русский язык: Теория: Учеб. Для 5-9 кл. обще-образоват. учеб. издание – 2-е изд. – М.: Просвещение, 1993. – 256с. (Дата обращения: 3.02.2013)
9. И.Наумова Понятие о приставке [Электронный ресурс] // goldrussian.ru [Сайт] URL:http://www.goldrussian.ru/ponjatie-o-pristavke-12.html (Дата обращения: 13.09.12)
10. И.Гаршин Русские префиксы [Электронный ресурс] // garshin.ru [Сайт] URL:http://www.garshin.ru/linguistics/model/systematics/russian/russian_prefixes.html (Дата обращения 13.09.12)
11. Суффиксы [Электронный ресурс] // Магия языка [Сайт] URL: http://wordsland.ru/magiclanguage/suff.html (Дата обращения 14.09.12)
12. Основные правила грамматики русского языка [Электронный ресурс] // priroda.inc.ru [Сайт] URL:http://priroda.inc.ru/blog/grammatika.html (Дата обращения 14.09.12)
13. Суффикс или окончание? [Электронный ресурс] : (Статья) / Л.С.Степанова // Первое сентября [Сайт]. (Дата публикации: №6 2001 газеты «Русский язык»). URL: http://rus.1september.ru/article.php?ID=200100608 (Дата обращения 04.02.2013).

14. Официальный сайт компании разработчика пакета программ Erwin [Электронный ресурс] // Erwin [Сайт] URL: http://erwin.com/ (Дата обращения 13.02.2013)

15. Методология функционального моделирования IDEF0 [Руководящий документ] // Госстандарт России (Издание официальное) – Москва, 2000 – 75с. (Дата обращения 13.02.2013)

16. MCR.DLL [Электронный ресурс] // http://macrocosm.narod.ru [Сайт] URL: http://macrocosm.narod.ru/index.html (Дата обращения 30.10.2012)

17. Официальный сайт разработчиков компании devart [Электронный ресурс] // Devart [Сайт] URL: http://www.devart.com (Дата обращения 13.02.2013)

18. Официальный сайт разработчика Delphi [Электронный ресурс] // Embarcadero [Сайт] URL: http://www.embarcadero.com/products/delphi (Дата обращения 13.02.2013)

19. Официальный сайт разработчика MySQL [Электронный ресурс] // MySQL [Сайт] URL: http://www.mysql.com/ (Дата обращения 13.02.2013)

20. Delphi 7 начинающему [Электронный учебник] // realcoding [Сайт] URL: http://www.realcoding.net/articles/predislovie-delphi-%E2%80%94-chto-eto.html (Дата обращения 18.03.13)

21. Что такое система управлениями базами данных MySQL [Электронный ресурс] // Веб-библиотека [Сайт] URL: http://www.weblibrary.biz/mysql/info/bazy-dannyh-sql (Дата обращения 17.02.13)

22. СУБД MySQL [Электронный учебник] // Hardline.ru [Сайт] URL: http://www.hardline.ru/2/15/747/#Doesl (Дата обращения 17.02.13)

23. Л.Г. Зубкова Принцип знака в системе языка // М.: Языки славянской культуры, 2010 − 752 с. − (Studia philologica) [Электронный ресурс] URL: http://philologos.narod.ru/ling/zubkova.htm (Дата обращения 5.05.2013)

Приложение А

(обязательное)

Руководство пользователя

Содержание

Введение

1 Область применения

2 Основные понятия и определения.

Общие сведения о программе

3 Назначение программы

4 Порядок загрузки, запуска и завершения программы

Описание функций

5 Описание функций доступных пользователю

5.1 Морфологический анализ

5.2 Разбор слова по составу

6 Описание функций доступных эксперту

Сообщения об ошибках

Введение

Данный документ является руководством пользователя для программы Морфемный анализатор v0.1.

1 Область применения

Областью применения программы является школы, ВУЗы и лица заинтересованные в изучении раздела морфемики русского языка.

2 Основные понятия и определения

Введем определения основных понятий использующихся в данном руководстве и значение кнопок приложения (Таблица А.1).

Морфемика − 1. Морфемный строй языка, совокупность вычленяемых в словах морфем и их типы. 2. Раздел языкознания, изучающий типы и структуру морфем, их отношения друг к другу и к слову.

Морфемный анализ (разбор слова по составу) – это действие, направленное на выделение в изучаемом слове его минимальных значимых частей (морфем).

Морфема − минимальная, не делимая на части, значимая часть слова.

Морфы – конкретные представители морфем в слове.

Корень слова — это обязательная часть слова, заключающая в себе основной компонент лексического значения слова.

Аффиксы — служебные морфемы, необязательные для слова, выражающие дополнительное лексическое или грамматическое значение слова.

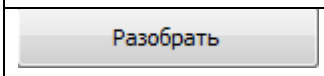
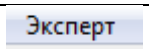
Префикс (франц. prefix) (приставка) − морфема (аффикс), стоящая перед корнем и изменяющая его лексическое или грамматическое значение.

Суффикс (лат. suffixus) − морфема (аффикс), следующая за корнем или основой и предшествующая окончанию.

Флексия (лат. flexio) (окончание) – морфема (аффикс), выражающая грамматические значения при словоизменении (склонении, спряжении).

Интерфикс (лат. interfixus) – морфема (аффикс), соединяющая корни или корень с суффиксом: например, русские -е-, -о- в сложных словах пароход, землетрясение.

Таблица А.1 – Кнопки

Иллюстрация кнопки	Название	Функция
	Разобрать слово по составу	Предназначена для запуска процедур разбора введенного пользователем слова.
	Раскрыть дерево разбора	Предназначена для отображения дерева всех возможных вариантов разбора слова.
	Сохранить дерево разбора	Сохраняет дерево всех возможных вариантов разбора слова в файл .txt. Имя файла совпадает с веденным словом.
	Открыть форму эксперта	Открывает форму эксперта для редактирования базы данных системы
Кнопки для редактирования базы данных морфем		
	First	Переход на первую запись в таблице
	Prior	Переход на предыдущую запись
	Next	Переход на следующую запись
	Last	Переход на последнюю запись
	Insert	Вставка новой записи перед текущей
	Delete	Удаление текущей записи с переходом на следующую

Продолжение таблицы А.1

Иллюстарция кнопки	Название	Функция
	Edit	Переводит источник данных в режим редактирования записи
	Post	Запись измененных данных из текущей записи в БД
	Cancel	Отмена изменений данных в текущей записи
	Refresh	Обновление данных в буфере источника
Кнопки для редактирования базы данных слов		
	First	Переход на первую запись в таблице
	Prior	Переход на предыдущую запись
	Next	Переход на следующую запись
	Last	Переход на последнюю запись
	Новое слово	Создает поля для ввода нового слова в базу данных
	Refresh	Обновление данных в буфере источника
	Редактировать	Переводит источник данных в режим редактирования записи
	Delete	Удаление текущей записи с переходом на следующую
	Save	Запись нового слова в БД

Общие сведения о программе

3 Назначение программы

Программа разработана для пользователей, целью которых является изучение принципов разбора слова по составу; морфем русского языка и орфографических правил, завязанных на морфемном анализе.

4 Порядок загрузки, запуска и завершения программы

Для загрузки приложения на свой компьютер скопируйте папку с программой, содержащей файл Морфемный анализ v0.1.exe, и БД morphemic_analysis.

Запустите программу двойным щелчком левой кнопки мыши по файлу Морфемный анализ v0.1.exe. В результате должно открыться окно программы (рисунок А.1). Программа готова к работе.

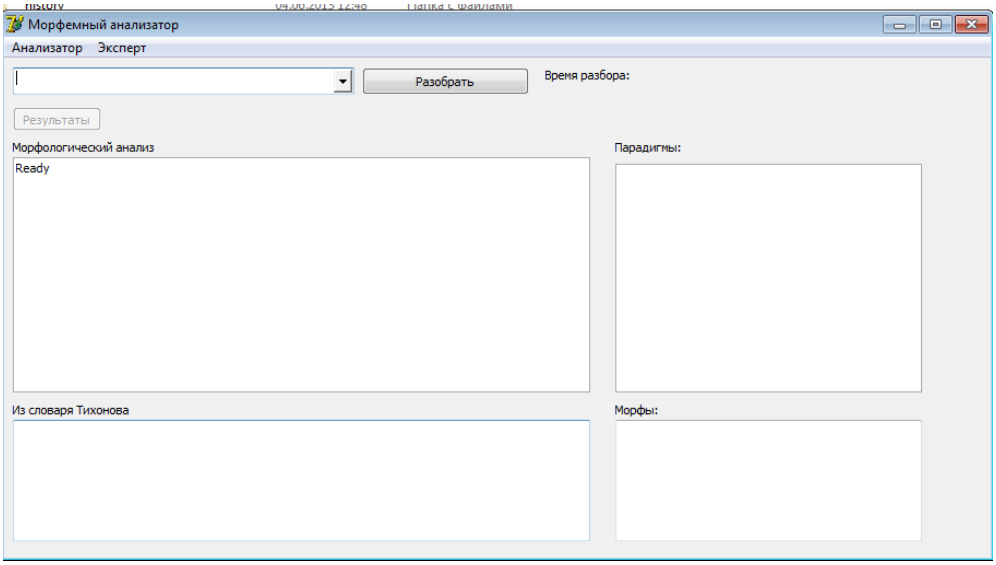


Рисунок А.1 – Стартовое окно программы

Для завершения работы приложения необходимо нажать кнопку закрытия приложения в правом верхнем углу окна.

Описание функций

5 Описание функций доступных пользователю

5.1 Разбор слова по составу

Для того чтоб разобрать слово по составу необходимо ввести его в поле ввода и нажать кнопку разобрать. Результат разбора отобразится в виде дерева разбора и в графическом виде в отдельном окне (рисунок А.3), морфологический разбор отобразиться в главном окне программы (Рисунок А.2).

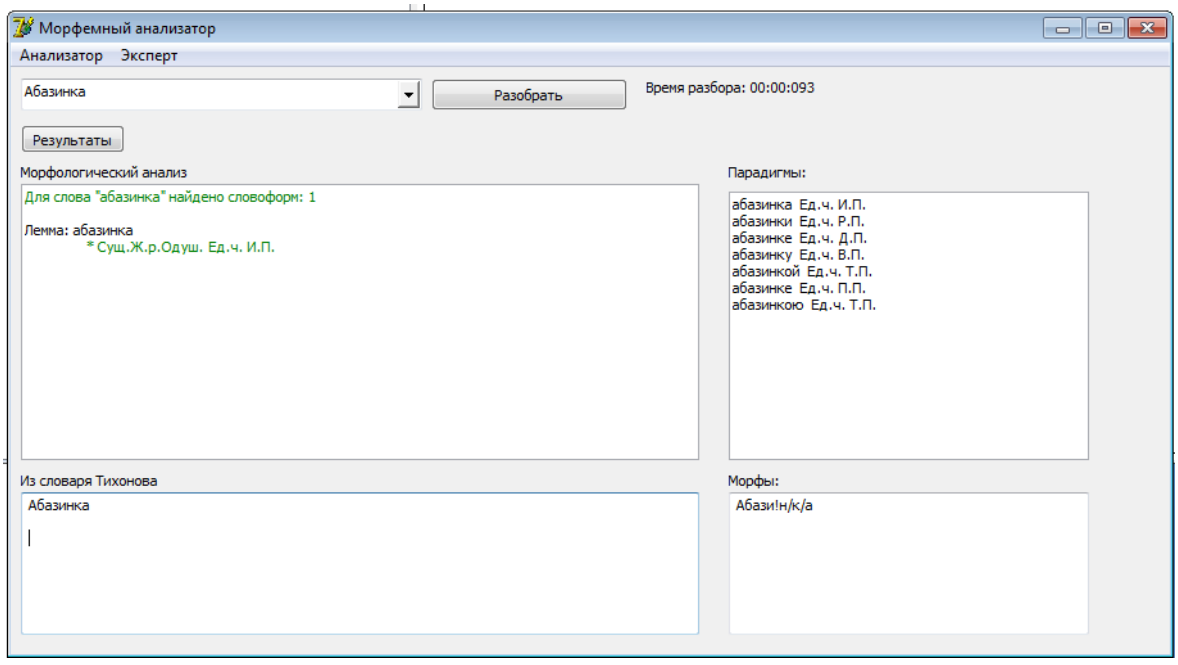


Рисунок А.2 – Результат разбора слова

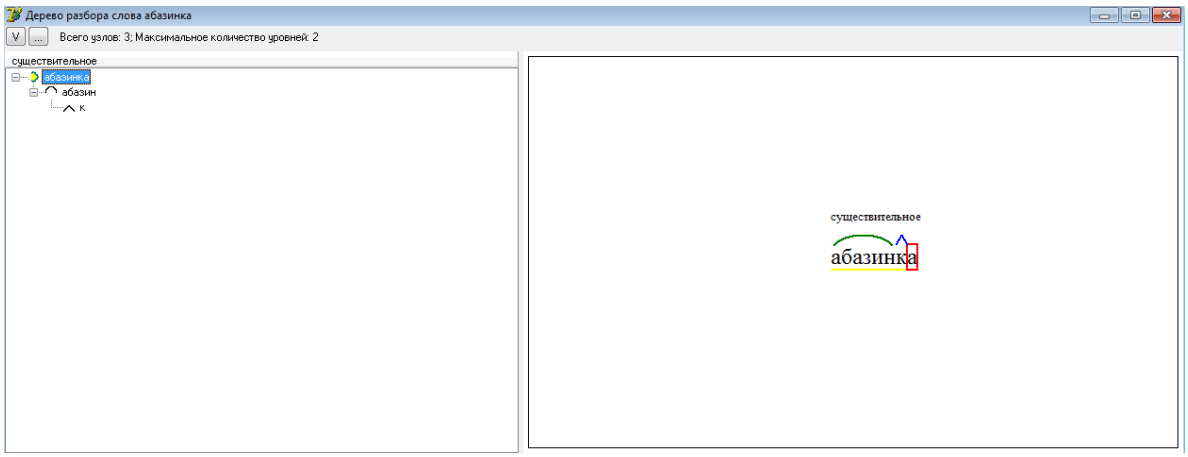


Рисунок А.3 – Результаты разбора слова

Для того чтобы просмотреть все дерево разбора необходимо нажать кнопку Раскрыть дерево разбора.

Справа от кнопки разобрать будет показано время, за которое был произведен разбор слова. Снизу отображаются варианты разбора слова, найденные в словаре А.Н.Тихонова.

Для того чтобы сохранить результаты разбора слова, нажмите кнопку Сохранить дерево разбора. В результате в папке с программой появится текстовый файл, имя файла будут совпадать с веденным для разбора словом плюс расширение .txt. Например, если вы ввели слов РАЗОБРАТЬ, то сохраненный файл будет РАЗОБРАТЬ.txt.

5.2 Морфологический анализ

В ходе морфемного разбора выполняется и морфологический анализ слова. Определяется часть речи, род, число, падеж, лемма и парадигма введенного слова. Результаты морфемного анализа отображаются в окнах сверху (рисунок А.2).

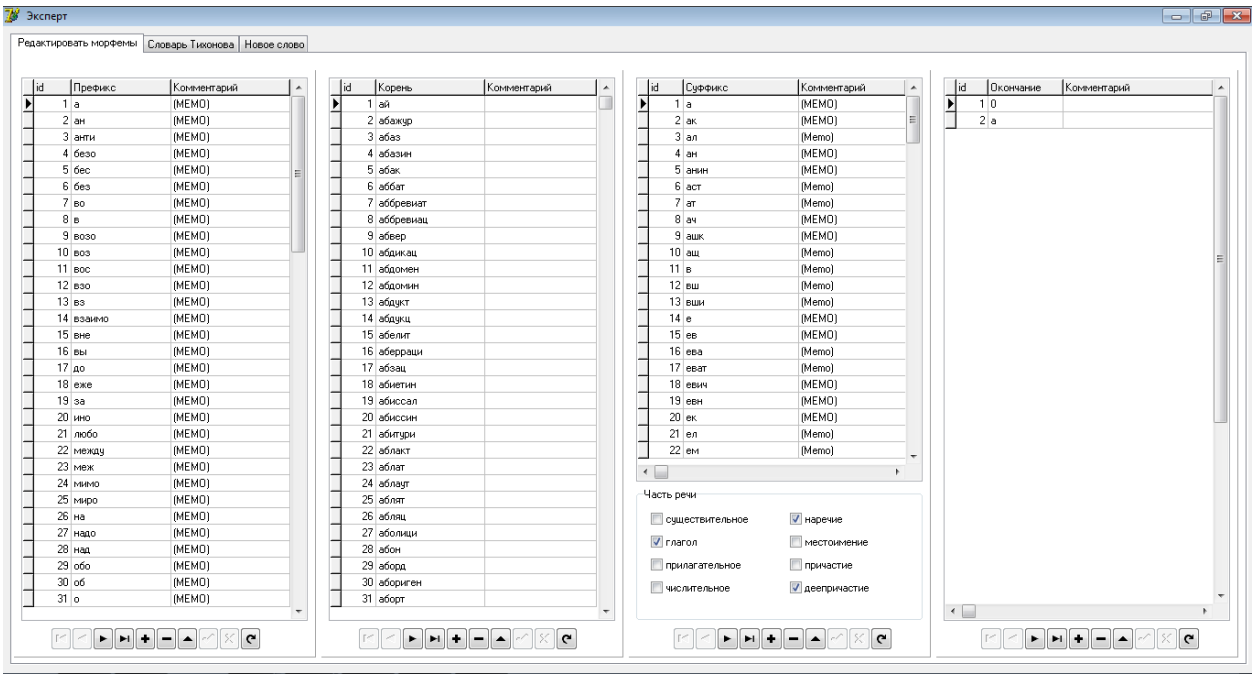
6 Описание функций доступных эксперту

Эксперту доступны функции редактирования таблиц морфем, добавления результатов разбора слов на основании словаря А.Н.Тихонова и добавление несловарных слов.

На рисунке А.4 представлено окно редактирования морфем базы данных. Для таблицы суффиксов снизу представлены данные о части речи, которым принадлежит тот или иной суффикс.

На рисунке А.5 представлена форма добавления слов в БД на основе словаря А.Н.Тихонова. В выпадающем списке предложен перечень морфем для выбранного слова, однако не запрещен и ручной ввод.

На рисунке А.6 представлена вкладка просмотра и редактирования слов и их разборов уже имеющихся в системе. Возможно изменение уже имеющихся слов и добавление новых слов (см. кнопки в таблице А.1).



РисунокА.5 – Редактирование морфем

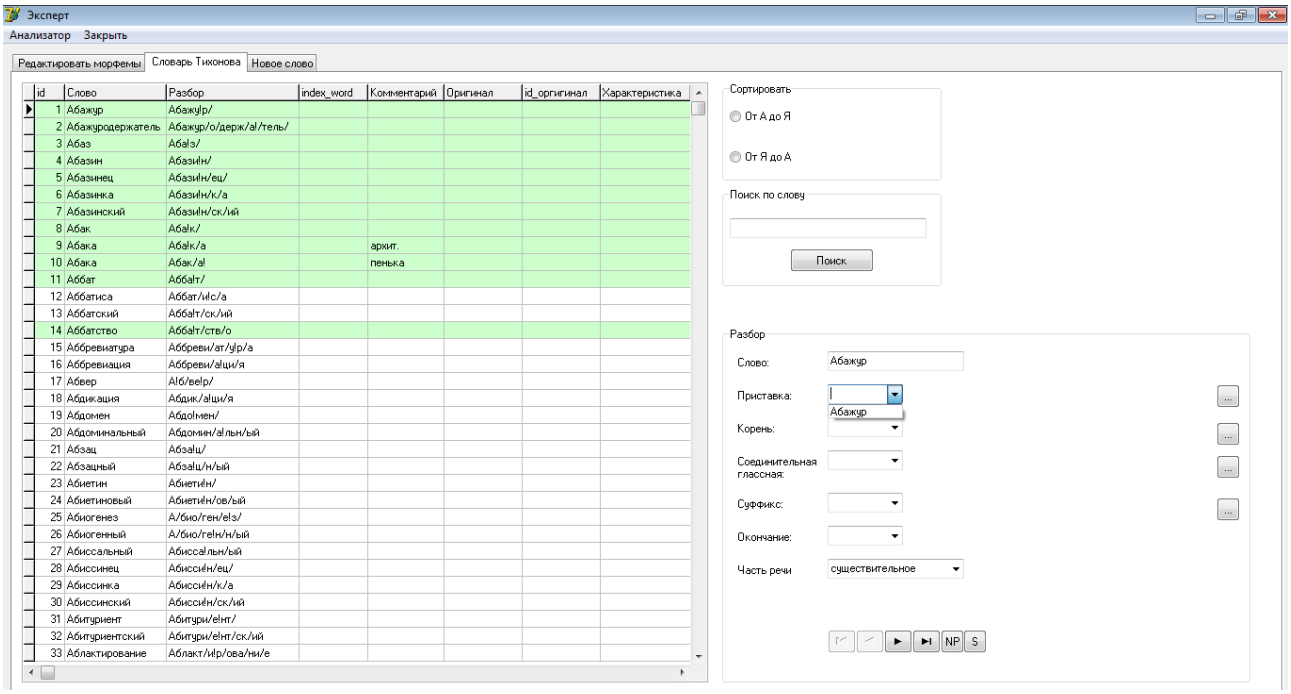
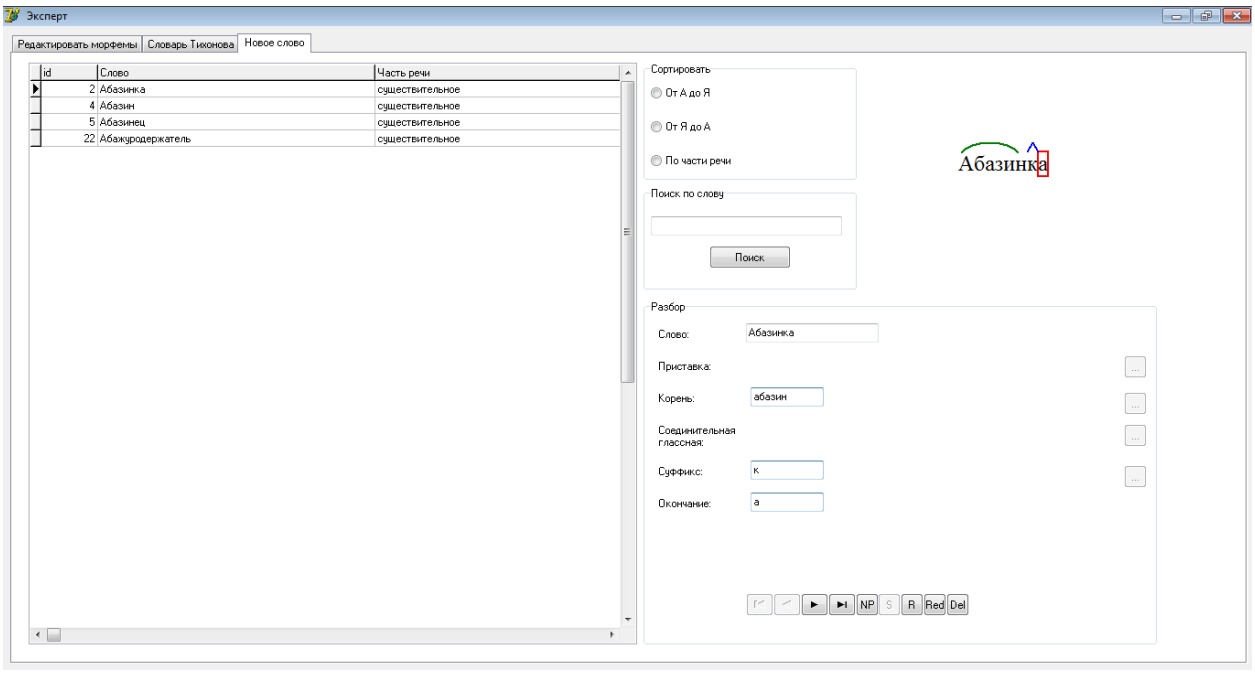


Рисунок А.6 – Добавление новых слов по словарю А.Н.Тихонова

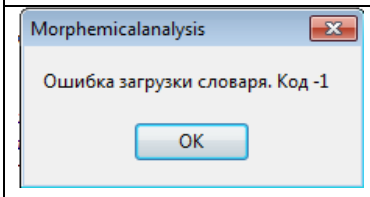
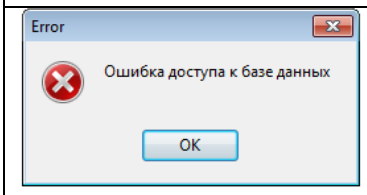
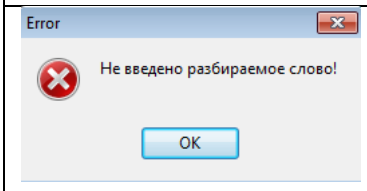
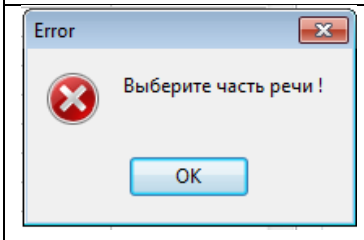


РисунокА.7 – Редактирование списка разобранных слов

Сообщения об ошибках

В таблице А.2 приведены типичные ошибки, с которыми можно столкнуться.

Таблица А.2 – Сообщения об ошибках

1. Иллюстрация сообщения об ошибке	2. Значение ошибки	3. Способ устранения
	Ошибка при загрузке библиотеки zal.mcr	Проверьте наличие файла zal.mcr в папке с приложением.
	Произошла ошибка обращения к базе данных morphemic_analysis.	
	Произошла ошибка при попытке сохранения нового слова в БД.	Введите значение разбираемого слова в соответствующее поле.
	Произошла ошибка при попытке сохранения нового слова в БД.	Выберите значение части речи для разбираемого слова.

Приложение Б

(обязательное)

Текст программы

Здесь должен быть текст программы 10 шрифтом, интервал 1, в одну колонку

Приложение В

(справочное)

Таблицы переменных и постоянных грамматических характеристик

Таблица постоянных грамматических характеристик

Таблица В.1 − Имя существительное

Cid	Описание
1	Существительное Мужского рода (неодушевленное)
2	Существительное Мужского рода (одушевленное)
3	Существительное Женского рода (неодушевленное)
4	Существительное Женского рода (одушевленное)
5	Существительное Среднего рода (неодушевленное)
6	Существительное Среднего рода (одушевленное)
7	Существительное Мужского-Женского рода (неодушевленное)
8	Существительное Мужского-Женского рода (одушевленное)
9	Существительное Мужского-Среднего рода (неодушевленное)
10	Существительное Мужского-Среднего рода (одушевленное)
11	Существительное Женского-Среднего рода (неодушевленное)
12	Существительное Женского-Среднего рода (одушевленное)
13	Существительное только множественное число(неодушевленное)
14	Существительное только множественное число (одушевленное)
15	Существительное *

Таблица В.2 − Имя прилагательное

Cid	Описание
20	Прилагательное
21	Местоимение
22	Местоименное прилагательное
23	Числительное собирательное
24	Числительное прилагательное
25	Числительное
26	Местоименное прилагательное (краткое)

В таблице нет специальных помет для безличных, многократных и вспомогательных глаголов.

Таблица В.3 − Спряжение

Cid	Описание
40	Глагол НСВ (несовершенного вида) невозвратный I спряжение
41	Глагол НСВ невозвратн II
42	Глагол НСВ возвратн I
43	Глагол НСВ возвратн II
44	Глагол СВ(совершенного вида) невозвратн I спряжение
45	Глагол СВ невозвратн II
46	Глагол СВ возвратн I
47	Глагол СВ возвратн II
48	Глагол СВ-НСВ I (двувидовый глагол)
48	Глагол СВ-НСВ II
50	Глагол (СВ)-НСВ возвратный I (совершенность носит потенциальный характер)
51	Глагол (СВ)-НСВ возвратный II

Таблица В.4 − Отглагольные формы

Cid	Описание
60	Причастие Настоящего времени (от НСВ I)
61	Причастие Настоящего времени (от НСВ II)
62	Причастие Настоящего времени (от НСВ I) страдательное значение(на -ся)
63	Причастие Настоящего времени (от НСВ II) страдательное значение
64	Причастие Прошедшего времени (от НСВ I)
65	Причастие Прошедшего времени (от НСВ II)
66	Причастие Прошедшего времени (от НСВ I) страдательное значение
67	Причастие Прошедшего времени (от НСВ II) страдательное значение
68	Причастие Прошедшего времени (от СВ I)
69	Причастие Прошедшего времени (от СВ II)
70	Причастие Прошедшего времени (от СВ I) страдательное значение
71	Причастие Прошедшего времени (от СВ II) страдательное значение
72	Страдательное Причастие настоящего времени (от НСВ I)
73	Страдательное Причастие настоящего времени (от НСВ II)
74	Страдательное Причастие прошедшего времени (от НСВ I)
75	Страдательное Причастие прошедшего времени (от НСВ II)
76	Страдательное Причастие прошедшего времени (от СВ I)
77	Страдательное Причастие прошедшего времени (от СВ II)

Продолжение таблицы В.4

Cid	Описание
78	Деепричастие (от НСВ I)
79	Деепричастие (от НСВ II)
80	Деепричастие (от СВ I)
81	Деепричастие (от СВ II)

Таблица В.5 − Остальные части речи

Cid	Описание
30	Наречие
31	Союз
32	Междометие
33	Частица
34	Предлог
35	Предикат
36	Вводное слово
37	Неизменяемое слово
200	Имя собственное *
201	Имя собственное мужского рода
202	Имя собственное женского рода
203	Отчество муж. род
204	Отчество женск. род
205	Фамилия
206	Название *
207	Географическое название
208	Географическое название мужского рода
209	Географическое название женского рода
210	Географическое название среднего рода
211	Географическое название только множественное число
212	Прилагательное образованное от геогр. названия
213	Аббревиатура
214	Аббревиатура (все прописные)
215	Аббревиатура (все ЗАГЛАВНЫЕ)
216	Сокращение кг,сек,см. Рисунок и.т п

Таблицы кодирования переменных грамматических характеристик

Таблица В.6 – Переменные грамматические характеристики существительного

Vid	Описание
0	Все формы одинаковы
1	Ед.ч. И.П. (единственное число, именительный падеж)
2	Ед.ч. Р.П.
3	Ед.ч. Д.П.
4	Ед.ч. В.П.
5	Ед.ч. Т.П.
6	Ед.ч. П.П.
7	Мн.ч. И.П. (множественное число, именительный падеж)
8	Мн.ч. Р.П.
9	Мн.ч. Д.П.
10	Мн.ч. В.П.
11	Мн.ч. Т.П.
12	Мн.ч. П.П.
13	только мн. ч. (все формы одинаковы)

Таблица В.7 – Переменные грамматические характеристики прилагательных и схожих частей

Vid	Описание
1	И.П. М.р ед.ч.од/неод (Именительный падеж, муж. род, ед. число, одушевленное и неодушевленное)
2	И.П. С.р ед.ч.од/неод
3	Р.П. М/С.р ед.ч.од/неод
4	Д.П. М/С.р ед.ч.од/неод
5	В.П. М.р ед.ч.неод.
6	В.П. М.р ед.ч.одуш.
7	В.П. С.р ед.ч.од/неод
8	Т.П. М/С.р ед.ч.од/неод
9	П.П. М/С ед.ч.од/неод
10	И. Ж.р ед.ч.од/неод
11	Р,Д,П и Ж ед.ч.од/неод
12	В. Ж.р ед.ч.од/неод
13	Т. Ж.р ед.ч.од/неод
14	И. Мн.ч. од/неод
15	Р.Мн.ч. од/неод

Продолжение таблицы В.7

Vid	Описание
16	Д. Мн.ч. од/неод
17	В. Мн.ч. неод.
18	В. Мн.ч. од.
19	Т. Мн.ч. од/неод
20	Т. Ж.р еч.од/неод
21	Кратк.форма М.р
22	Кратк.форма Ж.р
23	Кратк.форма С.р
24	Кратк.форма Мн. всех родов
25	Сравнительная степень
26	Сравнительная степень (параллельный вариант ее/ей)

Таблица В.8 – Переменные грамматические характеристики числительных

Vid	Описание	Vid	Описание
0	все формы одинаковы	12	М/С род В.П.
1	И.П.	13	М/С род В.П. одушевл
2	Р.П.	14	М/С род Т.П.
3	Д.П.	15	М/С род П.П.
4	В.П.	16	Ж род И.П.
5	В.П. одушевленное	17	Ж род Р.П.
6	Т.П.	18	Ж род Д.П.
7	П.П.	19	Ж род В.П.
8	Т.П. (параллельн)	20	Ж род В.П. одушевл
9	М/С род И.П.	21	Ж род Т.П.
10	М/С род Р.П.	22	Ж род П.П.
11	М/С род Д.П.		

Таблица В.9 – Переменные грамматические характеристики глагола

Vid	Описание	Vid	Описание
1	ИнФинитив	15	Повел. 1 лицо Мн.(ко многим)
2	Н.вр Ед.ч 1 лицо	16	Буд.вр Ед.ч 1 лицо
3	Н.вр Ед.ч 2 лицо.	17	Буд.вр Ед.ч 2 лицо
4	Н.вр Ед.ч 3 лицо	18	Буд.вр Ед.ч 3 лицо
5	Н.вр Мн.ч 1 лицо	19	Буд.вр Мн.ч 1 лицо
6	Н.вр Мн.ч 2 лицо	20	Буд.вр Мн.ч 2 лицо
7	Н.вр Мн.ч 3 лицо	21	Буд.вр Мн.ч 3 лицо
8	Пр.вр Ед.всех лиц М род	25	Н/Буд. вр Ед.ч 1 лицо
9	Пр.вр Ед.всех лиц Ж род	26	Н/Буд. вр Ед.ч 2 лицо
10	Пр.вр Ед.всех лиц С род	27	Н/Буд. вр Ед.ч 3 лицо
11	Пр.вр Мн.всех лиц родов	28	Н/Буд. вр Мн.ч 1 лицо
12	Повел. 2 лицо Ед.	29	Н/Буд. вр Мн.ч 2 лицо
13	Повел. 2 лицо Мн.	30	Н/Буд. вр Мн.ч 3 лицо
14	Повел. 1 лицо Мн.(к одному)		

Лингвисты как правило не различают времени у прилагательного, но раз такая информация была введена в словаре Зализнякак, то 2 характеристики имеют место.

Таблица В.10 – Переменные грамматические характеристики деепричастия

Vid	Описание
1	Настоящего времени
2	Прошедшего времен

Таблица В.11 – Прочее

Vid	Описание
0	NULL