МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Кафедра компьютерных систем

Садуахас Нургуль Сериковна

«Разработка и создание базы данных фонетического разбора слов русского языка»

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

по направлению высшего профессионального образования

230100.68 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Тема диссертации утверждена распоряжением по НГУ № 5 от «11» января 2012 г.

Тема диссертации скорректирована распоряжением по НГУ № 71 от «27» февраля 2013 г.

Руководитель

Барахнин Владимир Борисович

д.т.н, с.н с. ИВТ СО РАН

Новосибирск, 2013 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ…….3

ГЛАВА I…….....5-8

Подходы и технологии автоматизации анализа поэтических текстов……...

I.I Метр, ритм, фонетика……..5-6

I. II Фонетика и фонетический разбор слова……6-8

ГЛАВА II

ТЕХНИЧЕСКАЯ ЧАСТЬ ЗАДАЧИ……....9-18

II.I Выбор средств…….9

II.II Базы данных и информационные системы…….9-10

II. III Подпрограммы……...……..11-18

ЗАКЛЮЧЕНИЕ……19

Литература……...20

Приложение……21-25

ВВЕДЕНИЕ

В современном мире филология все более тесно взаимодействует с информатикой. Такое сближение было обусловлено двумя вещами: появлением мировой сети INTERNET и электронной полиграфии. Соответственно нам нужны методы и модели, позволяющие работать с новыми аналогами книг. В перспективе такие модели откроют широкие возможности для филологов. Автоматизация процесса обработки стихотворных текстов сделает возможным изучение больших поэтических и применение различных методов интеллектуального анализа данных.

Уровни структуры стиха, подобно уровням структуры произвольного сообщения, также представляют собой определенную иерархию. Для поэтического текста такими уровнями являются: ***метр, ритм, фонетика, лексика, грамматика, речевой жанр (композиционно-речевое целое), тематика, литературный жанр*** [1]**.** При этом процесс анализа стиха предусматривает первоначальное рассмотрение каждого уровня как самостоятельной смысловой единицы с последующим связыванием этих наблюдений с другими элементами структуры.

Актуальность. При автоматизации анализа различных уровней структуры стиха одной из важнейших задач является исследование его фонетической, метрической и ритмической структуры. Для этого необходимо иметь базу данных, содержащую фонетический разбор словоформ слов русского языка.

Цель исследования. Разработка структуры базы данных фонетического разбора словоформ слов русского языка и создание алгоритма ее наполнения.

Научная новизна. Впервые создается алгоритм автоматизированного наполнения базы данных фонетического разбора словоформ слов русского языка.

Объект и предмет исследования. Объектом исследования является данные содержащие в базе данных фонетики и словаре словоформ. Предметом исследования являются модели и алгоритмы автоматизированной генераций словоформ фонетического разбора слов русского языка.

Задачи работы состоит в следующем:

- изучение предметной области;
- исследование стилистических закономерностей поэтических материалов, в частности, исследование закономерностей образования стихотворных размеров;

- разработать и реализовать алгоритм обработки электронного словаря, извлекающий следующую информацию: орфографическая запись слова, ударение в слове, деление слова на слоги, фонетическая транскрипция слова и характеристика всех его звуков;
- разработать структуру базы данных фонетического разбора слов русского языка;
- для каждого извлеченного слова построить все его словоформы;
- создать удобный интерфейс для редактирования полученной информации.

ГЛАВА I

Подходы и технологии автоматизации анализа поэтических текстов

I. I Метр, ритм, фонетика

Анализ данного уровня стихов имеет весьма специфический характер, поскольку требует исследования фонетических характеристик лексем, каковое при анализе обычных сообщений почти никогда не проводится.

Сразу ответим на естественный вопрос: поскольку непосредственно в письменном сообщении его фонетические характеристики отсутствуют, можно ли отнести их к изложенной выше семиотической модели? Действительно, воспринять фонетические характеристики текста может лишь адресат информации: человек или запрограммированная на решение такой задачи информационная система, но ведь то же самое можно сказать и про семантические характеристики текста, например, смысл лексем. Здесь следует руководствоваться известным утверждением А. А. Ляпунова [2]: «информация всегда относительна, она зависит от того, какой информационной системой она воспринимается», на основании которого фонетические характеристики текста вполне могут быть отнесены к его синтаксическому уровню.

Анализ метра и ритма предполагает исследование чередования так называемых сильных и слабых звуков (несколько упрощенно – ударных и безударных слогов), при этом метр – «идеальная схема» чередования, а ритм – их реальное чередование, несколько отличающееся от идеального ввиду взаимодействия естественных свойств речевого материала и метрического закона[1].

Для такого анализа используются фонетические словари. Наиболее полным из известных нам сетевых фонетических словарей открытого доступа – «Словарь полного фонетического разбора» [3].

Однако использование этого словаря для анализа фонетически характеристик стиха осложняется тем, что в нем приведены только начальные формы слов, поэтому необходима генерация фонетической записи словоформ (сами словоформы содержатся в том или ином морфологическом словаре). Автоматизация этого процесса не совсем тривиальна, поскольку не существует строгих закономерностей расположения ударения в словоформах в зависимости от места его расположения в начальной форме слова.

При автоматическом анализе метра и ритма следует учитывать возможность использовать поэтом «нестандартных» ударений. Такая ситуация выявляется апостериори,

посредством сравнения соответствующей строки (использование в которой «правильного» ударения нарушает общий ритм) с соседними строками.

Фонетический анализ стиха включает исследование звуковых повторов и рифм (их типов, а также строфического строения стиха, составление словарей рифм и т.п.). Поскольку историческое развитие русской рифмы характеризуется снижением ее точности, постольку при автоматизированном анализе рифмы необходимо учитывать свойства фонем. Так, согласные фонемы различаются по месту образования, по способу образования по участию голоса и шума, по твердости и мягкости, по глухоте и звонкости. Некоторые из этих свойств для каждой фонемы каждого слова непосредственно указаны в словаре.

Разумеется, для анализа метрических и строфических характеристик стиха необходимы «эталонные» базы даны типичных размеров и строф.

I.II Фонетика и фонетический разбор слова

Фонетика – это раздел языкознания, который изучает звуковые единицы языка. Звуковой строй языка – особый ярус в структуре языка, а поэтому фонетика – самостоятельный раздел языкознания, который имеет свой особый предмет и задачи. В соответствии со структурой звуковой стороны языка фонетика изучает звуки, различные типы ударения и интонацию. Звуковая сторона – это необходимая форма существования слов, материальное выражение, без которого невозможно существование языка. Перед фонетикой ставятся следующие задачи:

- установить звуковой состав данного языка в определенный период его развития;
- изучить его в статическом состоянии или изучить эволюцию и развитие звуковой стороны на протяжении ряда эпох истории этого языка;
- определить последовательные изменения звуков речи и выяснить причины этих изменений;
- изучить фонетические явления данного языка в сравнении с фонетическими явлениями других родственных языков;
- исследовать звуковые структуры двух и более языков с целью нахождения у них общего и специфического.

Как и все лингвистические науки, фонетика может исследовать языковые явления в плане синхронии и диахронии.

Изучение фонетических явлений в плане синхронии – это исследование фонетики определенного языка в данный момент как готовой системы взаимосвязанных и взаимообусловленных элементов. Изучение фонетики в плане диахронии – это изучение фонетических явлений во времени, в изменении, в переходе одних явлений в другие.

Фонетический разбор

Фонетический разбор слова осуществляется по следующему плану:

1. Транскрибировать слово, поставив ударение.
2. Определить количество слогов, указать ударный.
3. Показать, какому звуку соответствует каждая буква. Определить количество букв и звуков.
4. В столбик выписать буквы слова, рядом – звуки, указать их соответствие.
5. Указать количество букв и звуков.
6. Охарактеризовать звуки по следующим параметрам:

гласный: ударный/безударный;

согласный: глухой/звонкий, твердый/мягкий.

Слог - минимальная произносительная единица речи, характеризующаяся максимальной слитностью своих компонентов. Количество слогов в слове определяется числом гласных звуков, т.к. именно гласный - вершина слога: о-го-род-ни-че-ство.

Ударение - это выделение с помощью фонетических средств одного из слогов слова. Ударный слог произносится длиннее, сильнее и отчетливее остальных. Словесное ударение - обязательный признак слова. Однако существует ряд слов, которые примыкают к другим словам и не несут на себе самостоятельного ударения (частицы, предлоги и некоторые др.).

Поскольку ударение определяет фонетическое слово, в отдельных случаях его границы могут не совпадать с морфологическим словом, например, перед экзаменом, мне больно, ранен был (два морфологических слова составляют одно фонетическое).

Ударение в слове одно, однако, если слово длинное, может появиться побочное ударение: электростанция.

Морфологический анализ – анализ текста, определяющий основные характеристики слова: нормальную форму слова – лемму, морфологическую часть речи, набор граммем – элементарных морфологических описателей, относящих словоформу к какому-то морфологическому классу. Например, число, род и падеж для существительного.

ГЛАВА II
ТЕХНИЧЕСКАЯ ЧАСТЬ ЗАДАЧИ

II. I Выбор средств

В связи с большими объемами извлекаемой информации для ее хранения и обработки удобно применять базу данных. Она имеет много преимуществ в сравнении с файловыми системами:

- простая организация данных
- возможность быстрой сортировки по любому критерию
- возможность одновременного доступа нескольких пользователей
- высокая скорость обработки данных

Для легкости управления БД, изменения и добавления данных существуют системы управления базами данных (**СУБД**). Одна из самых популярных СУБД в современных интернет – технологиях, бесспорно, **MySQL**[4]. К основным плюсам MySQL можно отнести его простоту, высокую скорость работы, быстроту обработки данных и оптимальную надежность.

Для корректной работы с MySQL логично использовать язык PHP, поскольку PHP — это открытый продукт, он обеспечен технической поддержкой талантливой команды разработчиков и проверен широкой аудиторией пользователей [5]. Кроме того, PHP может работать почти на любой операционной системе и с любым сервером. Практический характер PHP обусловлен пятью важными характеристиками:

- традиционность (многие конструкции языка позаимствованы из Си, Perl)
- простота
- эффективность
- безопасность
- гибкость

II. II Базы данных и информационные системы

В широком понимании под определение информационной системы (ИС) попадает любая система обработки информации.

Иногда используется более узкая трактовка понятия ИС как совокупности аппаратно-программных средств, включающая базы данных, системы управления базами данных (СУБД) и специализированные прикладные программы.

В любом случае основной задачей ИС является удовлетворение конкретных информационных потребностей в рамках конкретной предметной области.

Современные ИС немыслимы без использования баз данных и СУБД, поэтому термин «информационная система» на практике сливается по смыслу с термином «система баз данных».

База данных (БД) представляет собой совокупность специальным образом организованных данных, хранимых в памяти вычислительной системы и отображающих состояние объектов и их взаимосвязей в рассматриваемой предметной области.

База данных хранится и обрабатывается в вычислительной системе. Таким образом, любые внекомпьютерные хранилища информации (архивы, библиотеки и т.п.) базами данных не являются.

Данные в базе данных хорошо структурированы (систематизированы). Под структурированностью в данном случае понимается явное выделение составных частей (элементов), связей между ними, а также типизация элементов и связей, при которой с каждым типом элемента или связи соотносится определённая семантика и допустимые операции.

Структура базы данных обеспечивает эффективный поиск и обработку данных. Эффективность здесь главным образом определяется тем, как соотносятся гибкость и мощность возможностей (поиска и обработки) с затратами усилий и ресурсов.

Хранимые в базе данные имеют определенную логическую структуру – иными словами, описываются некоторой моделью представления данных (моделью данных), поддерживаемой СУБД. К числу классических относятся следующие модели данных: иерархическая, сетевая, реляционная. Кроме того, в последние годы появились и стали более активно внедряться на практике следующие модели данных: постреляционная, многомерная, объектно-ориентированная. Разрабатываются также всевозможные системы, основанные на других моделях данных, расширяющих известные модели. В их числе можно назвать семантические, концептуальные, ориентированные, объектно-реляционные и дедуктивно-объектно-ориентированные. Некоторые из этих моделей служат для интеграции баз данных, баз знаний и языков программирования. В некоторых СУБД поддерживаются одновременно несколько моделей данных.

Система управления базами данных (СУБД) – это комплекс языковых и программных средств, предназначенный для создания, ведения и совместного использования БД многими пользователями. Обычно СУБД различают по используемой модели данных.

Так, СУБД, основанные на использовании реляционной модели данных, называют реляционными СУБД.

Постановка задачи

На сегодняшний день в БД с помощью языка PHP посредством использования регулярных выражений занесены страницы сайта http://slovonline.ru/slovar_el_fonetic/. Создан понятный интерфейс для внесения корректировок в полученные данные. Кроме того, создана таблица поэтических размеров и разработан алгоритм анализа конкретного стихотворного текста. Пользователь может ввести стих, и эти стихотворные строки будут подвергнуты фонетическому разбору.

II. III Подпрограммы

Моя работа посвящена задаче: из сайта http://slovonline.ru/slovar_el_fonetic/ - «Словарь полного фонетического разбора» составить аналогическую базу данных фонетического разбора слов русского языка. Где каждому слову отдельно извлекаются следующие информации: орфографическая запись слова, ударение в слове, деление слова на слоги, фонетическая транскрипция слова и характеристика всех его звуков. Учитывая специфичность стиля и его преимущества, я использовала алгоритмы, учитывающие информацию о закономерностях текстовой структуры, например, общих для всех документов массива синтаксических и семантических конструкциях.

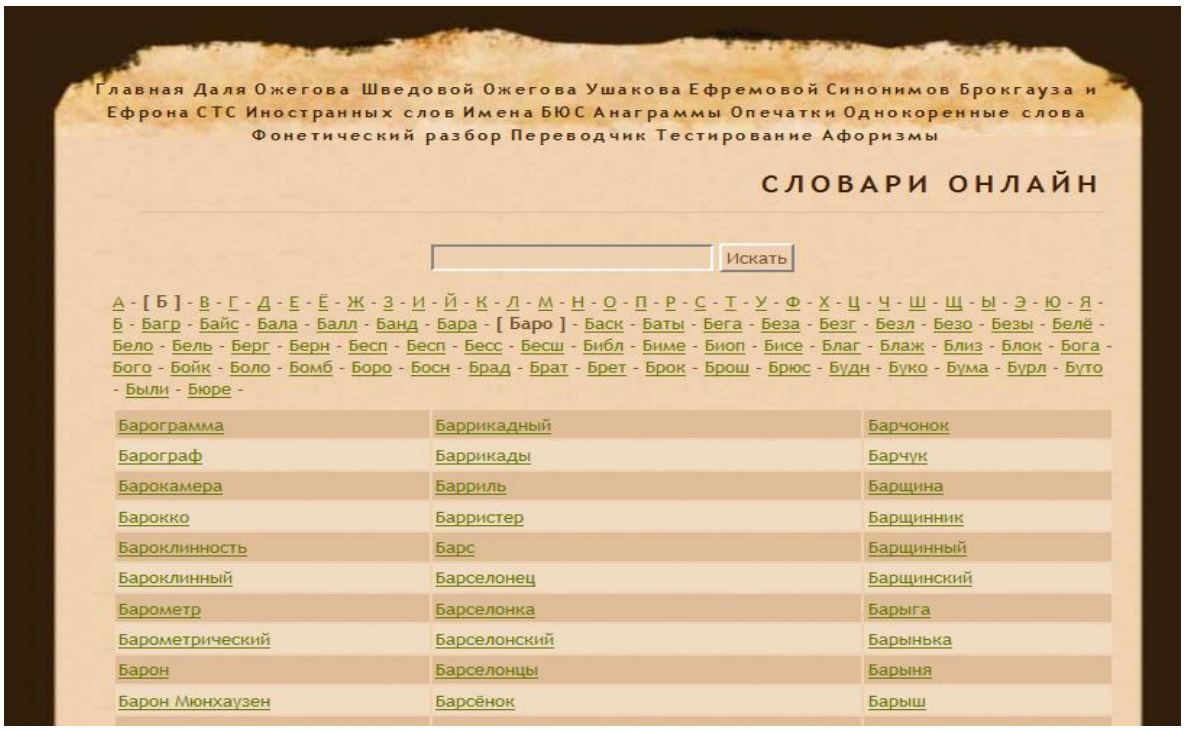


Рис 1. Главная страница «Словарь полного фонетического разбора»

http://slovonline.ru/slovar_el_fonetic/

Однако использование словаря «Словарь полного фонетического разбора» для анализа фонетически характеристик стиха осложняется тем, что в нем приведены только начальные формы слов, поэтому необходима генерация фонетической записи словоформ (сами словоформы содержатся в том или ином морфологическом словаре). Автоматизация этого процесса не совсем тривиальна, поскольку не существует строгих закономерностей расположения ударения в словоформах в зависимости от места его расположения в начальной форме слова.

Итак, подробней о самой программе. Она разбита на три части.

Первым шагом при разработке алгоритма анализа метрических и ритмических характеристик поэтических текстов является создание базы данных, содержащей фонетический анализ слов.

Процесс извлечения словаря начинается с создания в базе данных таблицы ссылок на разделы сайта, каждый из которых содержит слова на определенную букву алфавита. Инструментом поиска нужных строк, слов и наборов символов в тексте являются регулярные выражения. С их помощью извлекаются ссылки и заносятся в таблицу «letters» базы данных «dictionary». Поскольку слов много, каждый раздел разбит на

подразделы. Поэтому на следующем шаге создается таблица «subletters», содержащая ссылки на подразделы выбранной буквы.

Структура созданных таблиц сходна и имеет следующий вид:

Название поля	**Описание**	**Тип**
id	Идентификационный номер	INT(11)
letter	Название раздела / подраздела	VARCHAR(5)
refer	Ссылка на раздел/ ссылка на подраздел, относящийся к выбранному разделу	VARCHAR(50)

Таблица 1. «subletters» - содержащая ссылки на подразделы выбранной буквы

Тип связи между таблицами «letters» и «subletters» ─ «один ко многим», поскольку в каждом разделе есть несколько подразделов, но каждый подраздел принадлежит только одному разделу.

Следующий шаг ─ извлечение слов вместе с их фонетическим анализом из выбранного подраздела. Структура информации по каждому слову представлена на рисунке ниже, на примере слова «автомобиль»:

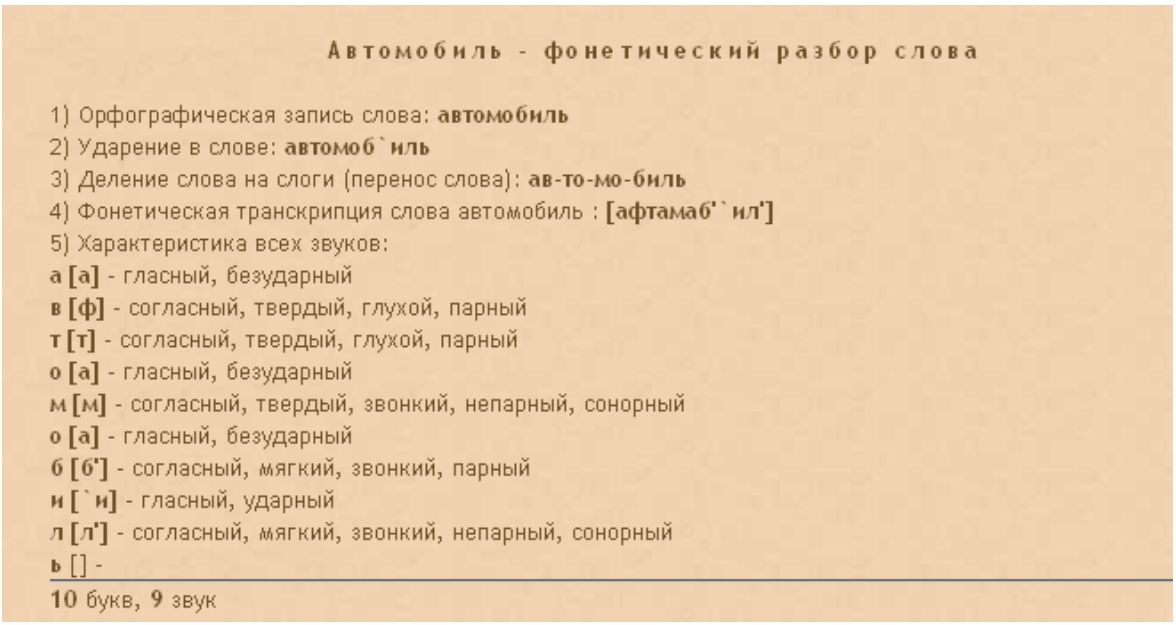


Рис 2. Структура слово «автомобиль»

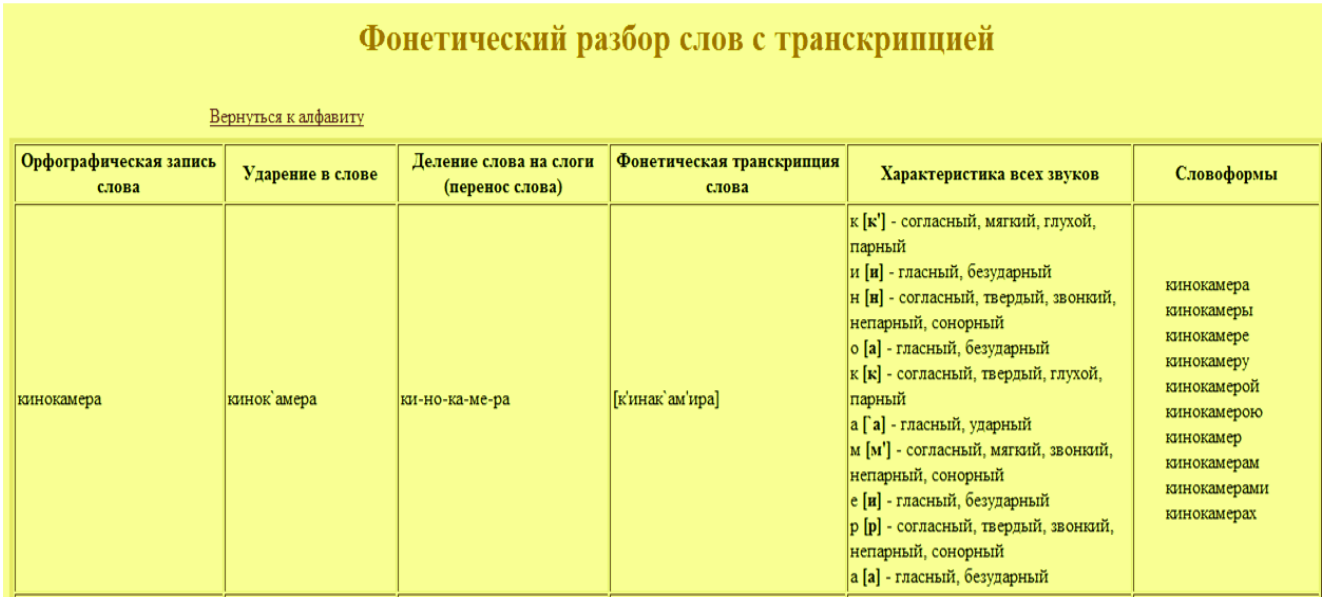
С помощью регулярных выражений со страницы извлекаются следующие элементы фонетического разбора: орфографическая запись слова, ударение, деление слова на слоги, фонетическая транскрипция и характеристика всех звуков. Результаты поиска и извлечения вносятся в созданную ранее базу данных и распределяются по соответствующим столбцам таблицы «phonetics». Структура таблицы:

Название поля	**Описание**	**Тип**
id	Идентификационный номер	INT(11)
spelling	Орфографическая запись	VARCHAR(100)
accent	Ударение в слове	VARCHAR(100)
syllables	Деление слова на слоги	VARCHAR(100)
transcription	Транскрипция слова	VARCHAR(100)

sounds	Характеристика звуков	VARCHAR(3000)
wordforms	Словоформы	VARCHAR(3000)

Связь этой таблицы и таблицы подразделов «subletters» принадлежит типу «один ко многим».

Для наглядности алгоритма извлечения словаря создан интерфейс, отражающий вид и содержимое таблицы базы данных.



Фонетический разбор слов с транскрипцией

Вернуться к алфавиту

Орфографическая запись слова	Ударение в слове	Деление слова на слоги (перенос слова)	Фонетическая транскрипция слова	Характеристика всех звуков	Словоформы
кинокамера	кинок`амера	ки-но-ка-ме-ра	[к'инак`ам'ира]	к [к'] - согласный, мягкий, глухой, парный и [и] - гласный, безударный н [н] - согласный, твердый, звонкий, непарный, сонорный о [а] - гласный, безударный к [к] - согласный, твердый, глухой, парный а [`а] - гласный, ударный м [м'] - согласный, мягкий, звонкий, непарный, сонорный е [и] - гласный, безударный р [р] - согласный, твердый, звонкий, непарный, сонорный а [а] - гласный, безударный	кинокамера кинокамеры кинокамере кинокамеру кинокамерой кинокамерою кинокамер кинокамерам кинокамерами кинокамерах

Рис 3. Собственно разбор слова (пример: слово «кинокамера»)

Следующий этап ─ разработка алгоритма автоматического образования фонетического анализа словоформ на основе фонетического разбора исходного слова. Для определения метра и ритма стихотворения в первую очередь необходимо получить базу ударений словоформ. Для этого можно воспользоваться данными интернет – проекта международной этимологической базы данных под названием "Вавилонская Башня". На страницах сайта размещен морфологический анализатор, содержащий среди прочего полные акцентуированные парадигмы для каждого слова, имеющегося в словаре программы. Запрос на обработку конкретного слова отсылается через форму, ниже которой появляется результат.

Рис 4. «Вавилонская Башня» международный Интернет – проект Сергея Анатольевича Старостина

В данной приложении мы имеем возможность ознакомиться с компьютерными базами данных по словарям Ожегова, Зализняка и Мюллера, а также проанализировать любое русское слово и получить его полную акцентуированную парадигму.

В базе данных каждое заглавное слово имеет отсылку к программе автоматического морфологического анализа. Эту программу можно вызвать и в качестве отдельного окна. В последнем случае введено может быть любое русское или английское слово в произвольной грамматической форме. Программой анализа выдаются следующие сведения:

Для русского слова –

- исходная словоформа (по А. А. Зализняку);
- словарная информация, то есть морфологический индекс русского слова и имеющиеся комментарии из Грамматического Словаря А. А. Зализняка;

- перевод, то есть набор словарных статей из словаря Мюллера, в которых содержится соответствующее русское слово, с готовыми отсылками на соответствующие словарные статьи;
- морфологическая характеристика введенного русского слова. В случае многозначности введенной формы выводятся все варианты анализа.

Затем приводятся полные акцентуированные парадигмы для каждого из результатов анализа.

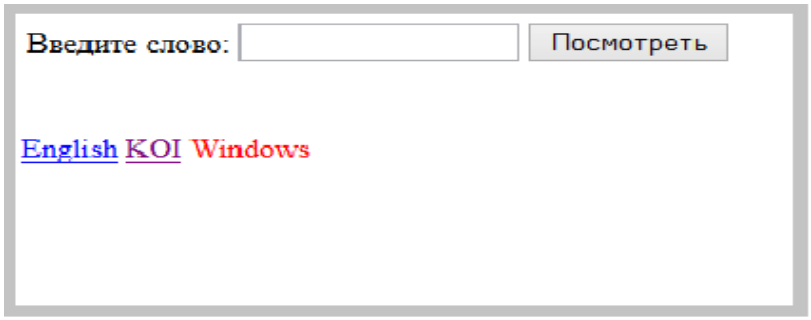


Рис 5. Морфологический анализ

1. Исходная форма: атлас

Словарная информация: м 1а (географич.)

Перевод: atlas I;atlas II;roadbook; satin;

Морфологическая характеристика: Ns ,Asi

2. Исходная форма: атлас

Словарная информация: м 1а, Р2 (ткань)

Морфологическая характеристика: Ns ,Asi

1 вариант:

	Ед. число	Множ. число
Именительный	а'тлас	а'тласы
Родительный	а'тласа	а'тласов
Дательный	а'тласу	а'тласам
Винительный неод.	а'тлас	а'тласы
Творительный	а'тласом	а'тласами
Предложный	а'тласе	а'тласах

2 вариант:

	Ед. число	Множ. число
Именительный	атла'с	атла'сы
Родительный	атла'са	атла'сов
Родительный 2	атла'су	
Дательный	атла'су	атла'сам
Винительный неод.	атла'с	атла'сы
Творительный	атла'сом	атла'сами
Предложный	атла'се	атла'сах

Рис 6. Пример: морфологический анализ слово «атлас»

Итак, для имеющихся в нашей базе данных слов поочередно автоматически делается запрос при помощи функции «fopen». Далее при помощи регулярных выражений из полученного ответа программы извлекаются словоформы вместе с ударениями.

Так обстоит дело с базой данных на сегодняшний день. Сейчас ведется работа над ее улучшением и дополнением, а также над доработкой интерфейса с целью получения возможности редактирования информации.

В области анализа стихотворений реализована предварительная обработка текста, которая включает в себя извлечение всех слов, приведение их к начальной форме, определение части речи и подсчет количества вхождений каждого слова независимо от его словоформ.

ЗАКЛЮЧЕНИЕ

- Изучена предметная область фонетического разбора слов русского языка;
- разработана структура базы данных фонетического разбора словоформ слов русского языка;
- создан алгоритм наполнения базы;
- разработан и реализован алгоритм обработки электронного словаря, извлекающий следующую информацию: орфографическая запись слова, ударение в слове, деление слова на слоги, фонетическая транскрипция слова и характеристика всех его звуков;
- для каждого извлеченного слова в словаре построены все его словоформы;
- создан понятный интерфейс для внесения корректировок в полученные данные.

В настоящей работе намечены основные подходы к автоматизации процесса статистического анализа низших структурных уровней (метр, ритм, фонетика) русских поэтических текстов. Результаты такого анализа позволят существенно расширить возможности филологов, исследующих как указанные уровни стихов, так и их семантические и прагматические характеристики, в том числе избавить филологов от рутинной работы, расширить круг анализируемых произведений, уменьшив зависимость качества сравнительного анализа от личной эрудиции исследователя, а также применять различные методы интеллектуального анализа данных.

Литература

1. Д. М. Магомедова. Филологический анализ лирического стихотворения/ Издательский центр «Академия», М. : 2004. – С. 4-5.
2. А. А. Ляпунов. О соотношении понятий материя, энергия и информация // В: А.А.Ляпунов. Проблемы теоретической и прикладной кибернетики. Новосибирск: Наука, 1980. С. 320-323.
3. Словарь полного фонетического разбора. http://slovonline.ru/slovar_el_fonetic/
4. Руководство MySQL (http://dev.mysql.com/doc/)
5. Руководство PHP (http://www.php.net/manual/en/)
6. Международный интернет – проект «Вавилонская Башня». (http://starling.rinet.ru/intrab/)

Приложение

```
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML 3.2//RU">
<HTML><HEAD>
<TITLE>Фонетический разбор слов с транскрипцией</TITLE>
<META HTTP-EQUIV="CONTENT-TYPE" CONTENT="text/html; charset=Windows-1251">
</HEAD>
<STYLE>
th {font-size: 16pt; color: #0000FF;}
body {color: #0000FF;}
</STYLE>
<BODY bgcolor="#00EEEE" link="#00EEEE" vlink="purple">
<p style="color:#0000FF;; font-weight: bold; font-size:30px; text-align:center;">Фонетический разбор слов с транскрипцией</p>
<?php
ini_set('memory_limit', '512M');
   $conn = NULL;
   $table = "allphonetics";

   function Connect_to_MySQL()   //соединение с БД и создание таблицы
   { #1
      $user = "root"; $pass = "pass";
      $db = "sample";
      global $table;
      global $conn;
      $conn = mysql_connect("localhost", $user, $pass);
      if (!$conn ) die("Нет соединения с MySQL");
            mysql_query("CREATE DATABASE IF NOT EXISTS $db") or die("Нельзя создать $db: ".mysql_error());
      mysql_select_db($db, $conn) or die("Нельзя открыть $db: ".mysql_error());
            mysql_query("DROP TABLE IF EXISTS $table");
            $query = "CREATE TABLE $table
            (        id INT NOT NULL AUTO_INCREMENT, PRIMARY KEY(id),
```

```
                    spelling VARCHAR(100),
        accent VARCHAR(100),
        syllables VARCHAR(100),
        transcription VARCHAR(100),
        sounds VARCHAR(2000)
      )";
            if (!mysql_query($query, $conn))
                    die ("Нельзя создать таблицу $table: ".mysql_error());
    } #1

    function Add_to_database(&$dberror, $arg1, $arg2, $arg3, $arg4, $arg5)
    { #2
      global $table;
      global $conn;
      $query = "INSERT INTO $table (spelling, accent, syllables, transcription, sounds)
                            VALUES ('$arg1','$arg2','$arg3','$arg4','$arg5')";
      if (!mysql_query($query, $conn))
      {
        $dberror = mysql_error();
        return false;
      }
      return true;
    } #2

    function Get_from_html($base_adr,$page)
    { #3
                            // Извлекаем ссылки на подразделы, относящиеся к
фиксированной букве
            $page_content = file_get_contents($base_adr.$page);

      preg_match_all("'\s+href\s*=\s*\"([^>\s\"]+)\"\s+class'si",$page_content,$sub_letter);
            $sub_letter=array_slice(array_unique($sub_letter[1]), 30); // здесь ссылки на под
разделы
```

```
array_unshift($sub_letter, $page."p-1/"); // плюс ссылка на нулевой подраздел

foreach ($sub_letter as $value)
{
	$page_content = file_get_contents($base_adr.$value);

preg_match_all("'\s+href\s*=\s*\"([^>\s\"]+)\"\s+title[^>]+class'si",$page_content,$word)
; //извлекли слова из подраздела
	foreach ($word[1] as $key)
	{
		preg_match("'id-(\d+)'",$key,$id);
		$norm[$id[1]]="$key";
	}
	unset($word);
	ksort($norm); //отсортировали слова по алфавитному порядку
	$slova=array();
	foreach($norm as $val)
		{
			$page_content = file_get_contents($base_adr.$val);
			preg_match_all("'[1-
4]\).+?<b>(.*?)</b>'si",$page_content,$temp);  // фонет. разбор: 1-4 пункты
			preg_match_all("'<B>([^\s-
])</b>(.*?)(?=<B>|<hr>)'s",$page_content,$temp1);  // 5-й пункт
			$slova[]=array($temp[1], $temp1[1], $temp1[2]);
		}
	unset ($norm, $temp, $temp1);

			// Вывод таблицы на экран
	foreach($slova as $slovo)
	{
		echo"
			<tr>
				<td>{$slovo[0][0]}</td>
				<td>{$slovo[0][1]}</td>
```

```
                                        <td>{$slovo[0][2]}</td>
                                        <td>{$slovo[0][3]}</td>
                                        <td>\n";
                                for($i=0; $i<count($slovo[2]); $i++)
                                {       echo"\t\t\t\t\t{$slovo[1][$i]}{$slovo[2][$i]}\n";
                                        $slovo[2][$i]=preg_replace("'<b>(.*?)</b>'",            "\$1",
$slovo[2][$i]);
                                        $slovo[2][$i]=preg_replace("'<br>'", "\n", $slovo[2][$i]);
                                        $slovo[2][$i]=preg_replace("' '", " ", $slovo[2][$i]);
                                        $zvuki=$zvuki.$slovo[1][$i].$slovo[2][$i];
                                }
                                echo"\t\t\t\t</td>
                                        </tr>";
                                $slovo[0][3]=mysql_real_escape_string($slovo[0][3]);
                                $zvuki=mysql_real_escape_string($zvuki);

                                        // Заполнение таблицы базы данных
                                $dberror = "";
                                if
(!Add_to_database($dberror,$slovo[0][0],$slovo[0][1],$slovo[0][2],$slovo[0][3],$zvuki))
                                {
                                        print "<p>Ошибка: $dberror</p><br>\n";
                                        break;
                                }
                                unset($zvuki);
                        }
                        unset($slova);
                }
                unset($sub_letter);
                return 0;
        } #3

        echo "<table bordercolor=brown border='4' cellspacing='1' cellpadding='2'>";
        echo   "
```

```
                    <tr>
                <th>Орфографическая запись слова</th>
                <th>Ударение в слове</th>
                <th>Деление слова на слоги (перенос слова)</th>
                <th>Фонетическая транскрипция слова</th>
                <th>Характеристика всех звуков</th>
                    </tr>";
        $base_adr = "http://slovonline.ru";
      $page = file_get_contents($base_adr."/slovar_el_fonetic/");
      preg_match_all("'\s+href\s*=\s*\"([^>\s\"]+)\"\s+class'si",$page,$letter);    // извлечение ссылок на разные буквы алфавита
        array_unshift($letter[1], "/slovar_el_fonetic/b-0/"); // добавили ссылку на букву А
      Connect_to_MySQL();
        foreach ($letter[1] as $let)
              Get_from_html($base_adr, $let);
        //Get_from_html($base_adr, $letter[1][1]);
      mysql_close($conn);
      echo "\n</table>\n";
  ?>
  </BODY>
  </HTML>
```