

На правах рукописи

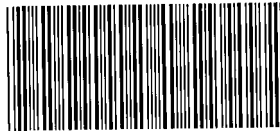


Килеев Вячеслав Васильевич

ЛИНГВИСТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ
АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ВЕРИФИКАЦИИ
ОРФОГРАФИИ И ГРАММАТИКИ ТЕКСТОВ
ФИННО-УГОРСКИХ ЯЗЫКОВ

05.13.12 – Системы автоматизации проектирования
(приборостроение)

Автореферат диссертации на соискание ученой степени
кандидата технических наук



005536019

31 ОКТ 2013

Санкт-Петербург – 2013

Работа выполнена на кафедре информационно-вычислительных систем Поволжского государственного технологического университета.

- Научный руководитель: доктор технических наук, профессор
Сидоркина Ирина Геннадьевна
- Официальные оппоненты: Гатчин Юрий Арменакович,
доктор технических наук, профессор, заведующий кафедрой «Проектирование и безопасность компьютерных систем» ФГБОУ ВПО «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»
- Донецкая Юлия Валерьевна,
кандидат технических наук, доцент, старший научный сотрудник ОАО «Концерн «ЦНИИ «Электроприбор»
- Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Марийский государственный университет»

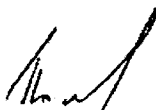
Защита состоится «20» ноября 2013 года в 15-50 на заседании диссертационного совета Д 212.227.05 Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики по адресу: 197101, Санкт-Петербург, Кронверкский пр., д. 49.

Отзывы на автореферат, заверенные печатью, просим направлять по адресу: 197101, Санкт-Петербург, Кронверкский пр., д. 49, СПб НИУ ИТМО, ученому секретарю диссертационного совета Д 212.227.05.

С диссертацией можно ознакомиться в библиотеке Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики.

Автореферат разослан «20» октября 2013 года.

Ученый секретарь
диссертационного совета Д 212.227.05
кандидат технических наук, доцент



В. И. Поляков

ОБЩАЯ ХАРАКТЕРИСТИКА ДИССЕРТАЦИИ

Актуальность темы диссертационного исследования

Представление текста на естественном языке согласно литературной норме является естественной характеристикой автоматизированных систем работы с текстом, в том числе и САПР. Результаты исследований показали, что научные и учебные источники содержат примерно 0,2% неправильно написанных слов. Среди всех неправильно написанных слов в тексте 90% составляют опечатки, которые могут быть обнаружены и устранены компьютерной или автоматизированной системой верификации орфографии, остальные 10% требуют верификации грамматики.

Актуальным является вопрос разработки и исследования автоматизированных систем верификации орфографии и грамматики для языков с малым количеством носителей языка. Системы проверки орфографии и грамматики разрабатывались в основном для языков, для которых имеются лингвистические корпуса (английский, финский, русский и др.). Кроме того, эти системы реализованы либо только как стационарные информационные приложения, либо как дополнительная функция другого приложения (обычно текстового процессора). Реализация системы в виде веб-приложения расширяет функционал автоматизированной системы.

Сегодня исследования в направлении реализации языковых возможностей автоматизированных систем ведут такие ученые, как норвежский лингвист Т. Тростеруд (финский, саамский языки), И.С. Ашманов (русский язык). В то же время работы по созданию систем верификации орфографии и грамматики для языков с малым количеством носителей, таких как марийский, удмуртский, мордовский и др., являются важными и актуальными. В настоящее время известны различные методы верификации орфографии и грамматики, при этом для разных групп языков применяются разные методы. В этой области можно выделить работы следующих ученых-лингвистов: Й. Луутонен, К. Н. Сануков (марийский язык), К. Коскенниemi (финский язык). Также можно назвать работы, выполняемые по проекту «Hunspell» (Венгрия). Вопросы распознавания языка представлены в трудах Х. Зянга, А. Г. Коробейникова (вьетнамский язык). Следует заметить, что исследования лингвистов без использования компьютерных систем малоэффективны.

Для построения систем проверки орфографии и грамматики языков, в том числе и финно-угорских, не существует больших лингвистических корпусов размером порядка 1 млрд словоупотреблений. Поэтому возникает необходимость исследования и разработки возможности применения для них методов, использующих алгоритмы, для которых все лингвистические данные языка предварительно задаются в специальных лингвистических конструкциях. Это алгоритмы стемминга. Также необходимо обратить внимание на реализацию морфоанализатора таких систем.

Перечисленные направления и задачи исследования являются актуальными для решения вопросов разработки алгоритмов, методов, архитектуры для систем верификации орфографии и грамматики финно-угорских языков и моделей

представления лингвистических конструкций для языков с отсутствующими лингвистическими корпусами.

Цель и задачи исследования

Целью настоящей работы является исследование и разработка программного и лингвистического обеспечения автоматизированной системы верификации орфографии и грамматики языков финно-угорской группы.

Основные задачи данной работы:

- анализ лингвистических компонент языков финно-угорской группы;
- анализ и классификация методов автоматизированной верификации орфографии и грамматики текстов на языках, имеющих и не имеющих лингвистические корпуса;
- анализ и классификация алгоритмов проверки орфографии для осуществления верификации орфографии;
- разработка алгоритма стемминга для работы с неограниченно длинными последовательностями аффиксов и для работы с частицами наравне с аффиксами;
- разработка модели распознавания предложения исходного языка по правилам грамматики, вводимым лингвистами, для естественных языков;
- разработка и исследование алгоритма генерации текстовых подсказок для правильных вариантов написания слова естественного языка в автоматизированных системах верификации орфографии текстов;
- разработка архитектуры и структуры программного обеспечения системы верификации орфографии и грамматики текстов финно-угорской группы;
- анализ особенностей лингвистического и программного обеспечения системы верификации орфографии и грамматики, реализующей человеко-машинный интерфейс через веб-приложение.

Объект исследования – программное и лингвистическое обеспечение автоматизированной системы верификации орфографии и грамматики текстов на языках финно-угорской группы.

Предметом исследования являются методы и алгоритмы верификации орфографии и грамматики текстов финно-угорских языков в автоматизированной системе.

Методы исследования базируются на использовании теории множеств, теории алгоритмов, теории принятия решений и методов объектно-ориентированного программирования, теории автоматизированного проектирования, методов системного анализа и формальных грамматик.

Научная новизна

- Предложены две новые лингвистические компоненты, отличающиеся от существующих функциональным назначением: инфлексии для согласования аффиксов и стема при генерации словоформ и параметры VARS (набор атрибутов) для согласования аффиксов между собой в длинных последовательностях.
- Разработан алгоритм стемминга, отличающийся от существующих использованием предложенных лингвистических компонент – инфлексии и параметров VARS – с циклической обработкой последовательностей аффиксов,

благодаря чему обеспечивается работа с неограниченно длинными последовательностями аффиксов и сокращается количество их повторений для высокоагглютинативных языков, а также отличающийся использованием различных способов написания групп аффиксов, благодаря которым возможно работать с частицами наравне с аффиксами, что позволяет проверять соответствие частиц словам.

- Предложена модификация алгоритма Дамерау-Левенштейна, которая благодаря вычислению длины каждого символа позволяет корректно рассчитывать меру разницы двух строк, представленных кодировкой с переменной длиной символов.

- Предложена модель распознавания предложения исходного языка по правилам грамматики, вводимым лингвистами, отличающаяся от существующих тем, что лексемы группируются в токены с атрибутами, хранящими семантику лексемы. Это позволяет сократить начальный алфавит грамматики и сделать правила вывода более наглядными.

- Разработан алгоритм генерации текстовых подсказок для правильных вариантов написания слова естественного языка, отличающийся от существующих обработкой введенных специальных параметров VARS, хранящих лингвистические характеристики слова.

- Предложена архитектура автоматизированной системы верификации орфографии и грамматики текста, которая, в отличие от существующих, благодаря выделению подсистемы верификации орфографии и подсистемы верификации грамматики позволяет осуществлять распараллеливание процесса верификации текста большого размера.

Основные положения, выносимые на защиту

- Лингвистические компоненты – инфлексии, параметры VARS.
- Алгоритм стемминга, позволяющий работать с частицами наравне с аффиксами и с неограниченно длинными последовательностями аффиксов.
- Модификация алгоритма Дамерау-Левенштейна.
- Модель распознавания предложения исходного языка по правилам грамматики, вводимым лингвистами.
- Алгоритм генерации текстовых подсказок для правильных вариантов написания слова.

Практическая значимость работы

Программная реализация автоматизированной системы проверки орфографии и грамматики финно-угорских языков, позволяющая осуществлять проверку текста на наличие орфографических и грамматических ошибок, имеет следующие преимущества:

- а) возможность верификации текста в среде Интернет;
- б) генерация подсказок по каждому варианту исправления неправильно написанного слова.

Программное обеспечение зарегистрировано в Реестре программ для ЭВМ Федеральной службы по интеллектуальной собственности, патентам и товарным знакам (свидетельство № 2013615729 от 19 июня 2013 г.).

Апробация результатов работы

Основные положения и результаты диссертационной работы докладывались и обсуждались:

- на научно-технической конференции «Исследования. Технологии. Инновации», 22-25 марта 2011 г., Йошкар-Ола;
- всероссийской научно-практической конференции «Информационные технологии в профессиональной деятельности и научной работе», 22-23 апреля 2011 г., Йошкар-Ола;
- конгрессе по интеллектуальным системам и информационным технологиям «IS&IT'11», 2-9 сентября 2011 г., Дивноморское, Россия;
- программе «Участник молодежного научно-инновационного конкурса» («У.М.Н.И.К.»), Йошкар-Ола;
- Первом Всероссийском фестивале науки в Республике Марий Эл, 7-9 октября 2011 г., Йошкар-Ола;
- Йо Форуме «Форум твоих идей», 20 ноября 2011 г., Йошкар-Ола;
- пятнадцатых Вавиловских чтениях «Инновационные ресурсы и национальная безопасность в эпоху глобальных трансформаций», 8-9 декабря 2011 г., Йошкар-Ола;
- XXVIII International Finno-Ugrist Students' Conference Tartu, 8-11 мая 2012 г., Тарту, Эстония;
- международной конференции «Автоматизация управления и интеллектуальные системы и среды», 9-15 октября 2012 г., Махачкала;
- конгрессе по интеллектуальным системам и информационным технологиям «IS&IT'13», 2-9 сентября 2013, Дивноморское, Россия.

Апробация и внедрение результатов диссертационной работы были проведены в ООО «ПешСайСофт», СГАУ РМЭ «Марийская база авиационной охраны лесов «Авиалесоохрана», ФГБОУ ВПО «ПГТУ», ФГБОУ ВПО «ЧГУ им. И.Н. Ульянова», ФГБОУ ВПО «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им В.И. Ульянова (Ленина)».

Результаты диссертации использовались в проектно-конструкторской деятельности ФГБОУ ВПО «ПГТУ» при подготовке и проведении Международной интернет-олимпиады по информатике и программированию, НИР 12.17/12 (гос. контракт 12.741.11.0050 от 27 апреля 2012 г.).

Работа выполнена при поддержке программы ФСР МФП НТС «Участник молодежного научно-инновационного конкурса 2012» («У.М.Н.И.К.») № 10508р/16915 от 1 июня 2012 г.

Публикации

По материалам диссертации опубликовано 15 печатных работ, в том числе три – в рецензируемых журналах, включенных в перечень ВАК.

Структура и объем работы

Диссертационная работа состоит из введения, четырех глав с выводами, заключения, списка использованной литературы (114 наименований). Общий объем 121 страница машинописного текста. Диссертация содержит 50 рисунков и 7 таблиц.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении рассмотрено современное состояние в области работы автоматизированных систем верификации орфографии и грамматики языков с отсутствующими лингвистическими корпусами, обоснована актуальность темы диссертации, сформулированы цель, задачи, определены объект и предмет исследования, научная и практическая значимость полученных результатов, приведена краткая характеристика основных разделов работы.

В первой главе исследуется общая схема работы систем верификации орфографии и грамматики, выделяются общие моменты в работе таких систем. Доказано, что основное отличие систем верификации орфографии и грамматики заключается в применяемых в них методах и алгоритмах. Выделены три категории методов: методы верификации орфографии, методы верификации грамматики и методы ранжирования списка вариантов исправления (англ. suggestion list). Для каждой из трех категорий приводится классификация и пояснение каждого из методов. Фрагмент классификации методов верификации орфографии и грамматики представлен на рис. 1, 2.

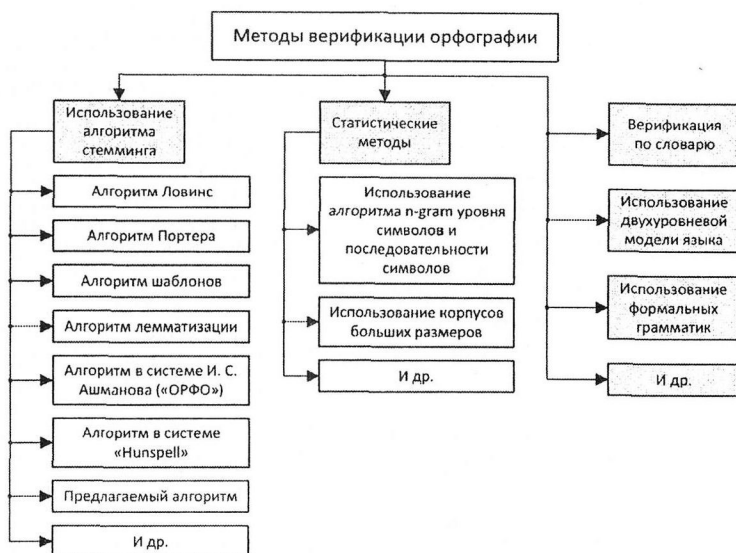


Рис. 1. Фрагмент классификации методов верификации орфографии

Среди методов ранжирования списка вариантов исправления ошибки рассматриваются следующие методы: ранжирования в алфавитном порядке, на основе вычисления редакторского расстояния (расстояние Хамминга, расстояние Левенштейна, расстояние Дамерау, расстояние Дамерау-Левенштейна), с использованием статистики по POS n-gram, с применением адаптивного алго-

ритма, на основе вычисления частоты использования предлагаемых вариантов и метод шаблонов.

Далее проведен анализ особенностей применения методов и алгоритмов в компьютерных системах, ориентированных на языки, для которых отсутствуют лингвистические корпуса. Для многих статистических методов требуется большое количество текстов, написанных на языке. Данные тексты получают из лингвистических корпусов. Лингвистический корпус – это филологически-компетентный массив языковых данных (как правило, множество текстов), отобранных в соответствии с некоторой исследовательской задачей и специально подготовленных, размеченных, структурированных, представленных в унифицированном виде.



Рис. 2. Фрагмент классификации методов верификации грамматики

При разработке системы верификации орфографии и грамматики текстов финно-угорских языков предложено исходить из того, что разрабатываемая система должна быть универсальной к языкам данной группы, т.е. ее можно было бы использовать для всех финно-угорских языков (как для языков с суще-

ствующими лингвистическими корпусами, так и для языков, для которых лингвистические корпуса отсутствуют). Так как для восточных финно-угорских языков таких корпусов нет, предложено использовать методы, для работы которых корпуса не требуются. Среди методов верификации орфографии обоснован выбор метода, реализующего алгоритм стемминга, метода верификации грамматики, основанного на использовании контекстно свободных формальных грамматик, и метода ранжирования списка вариантов исправления на основе вычисления меры редакторского расстояния.

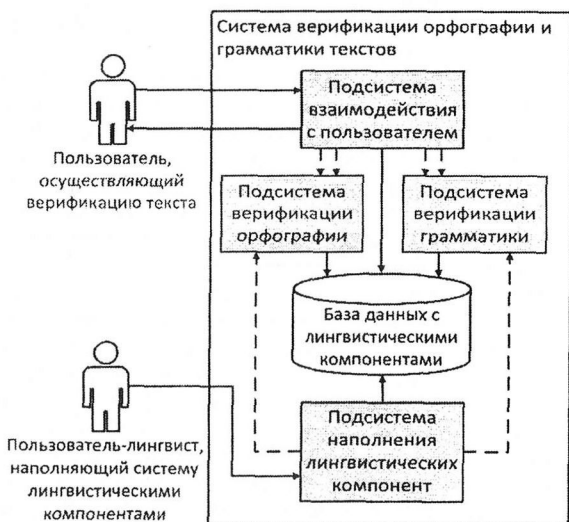


Рис. 3. Архитектура разработанной онлайн-системы верификации орфографии и грамматики

Предложенная архитектура онлайн-системы верификации орфографии и грамматики текста (рис. 3) отличается тем, что она является параллельной и позволяет осуществлять распараллеливание процесса верификации текста большого размера для языков финно-угорской группы.

Во второй главе приводится унифицированное обозначение лингвистических компонент, применяемых в системах верификации орфографии и грамматики. Основными лингвистическими компонентами разработанной системы являются стемы и аффиксы. Стем – неизменяемая часть слова, которая не обязательно совпадает с морфологическим корнем слова. Аффикс – изменяемая часть слова, при помощи которой образуется новая словоформа. Среди других компонент, применяемых в существующих системах верификации орфографии и грамматики, выделим также группы аффиксов, последовательности аффиксов, *n-gram*, правила вывода формальной грамматики, терминальные и нетерминальные символы формальной грамматики.

Для описания работы предложенного алгоритма стемминга и записи правил грамматики в подсистеме верификации грамматики введены две новые лингвистические компоненты: инфлектионы (англ. to inflect – изменять окончание слова) и параметры VARS. Инфлектион – компонента, обозначающая лингвистические характеристики, от которых зависит словообразование, и используемая для согласования стема и прикрепляемых к нему аффиксов. Параметры VARS – лингвистические характеристики, используемые для согласования нескольких аффиксов в длинной цепочке (два и более аффикса), а также для генерации текстовых подсказок о словоформе.

Далее в главе рассматривается предложенная модель распознавания предложения исходного языка, используемая в подсистеме верификации грамматики. Для того чтобы было меньше путаницы в терминах, будем называть систему верификации грамматики также системой верификации предложения.

Правила языка, вводимые лингвистами, представляют собой систему, которая является контекстно свободной грамматикой по классификации Хомского (тип 2, нумерация с нуля). Предварительно предложение исходного языка проходит лексический анализ в подсистеме верификации орфографии, в которой определяется принадлежность слова (лексемы) некоторому классу – токenu. Выделяются классы, соответствующие частям речи: существительное, прилагательное, числительное, глагол и т.д. Все эти классы получают из подсистемы верификации орфографии. Каждая лексема включает в себя дополнительные атрибуты, которые передаются для синтаксического анализа в подсистему верификации грамматики. Такими атрибутами являются условное обозначение цепочки аффиксов, по которой была сгенерирована лексема, и параметры VARS. У существительных, например, атрибуты хранят падеж, число и лицо. У глагола появляются такие атрибуты, как время, число, лицо.

Изначально все предложение обрабатывается лексическим анализатором, и синтаксический анализатор получает уже набор токенов с соответствующими атрибутами. Задача синтаксического анализатора – проверить правильность построения конкретного предложения на основе тех правил, которые ввели лингвисты. Сложность заключается в следующем: предложение может быть построено различными способами, и конкретная часть речи может находиться в различных местах предложения. Поэтому существуют различные формы представления одного предложения, а значит, различные лингвистические правила. Первая задача – определить по первому токenu, какое правило ему более подходит, затем попытаться свернуть этот токен, возможно, захватив следующий токен в некоторую лингвистическую конструкцию, т.е. некоторый нетерминал или правило. Далее предложение разбирается слева направо. Каждый раз выбирается некоторая основа, которую непосредственно можно свернуть к нетерминалу. Процесс продолжается до тех пор, пока все предложение не свернется к нетерминалу, который называется «предложение_языка». Получение нетерминала «предложение_языка» означает, что получена правильная структура предложения. Если уже достигнута «предложение_языка», но в исходном предложении еще остались несвернутые лексемы, то это, возможно, ошибка в предложении. Также это может свидетельствовать

о том, что лингвистами введены не все правила, необходимые для верификации предложения. Третий вариант, когда предложение рассмотрено до точки, но предложение не свернулось до нетерминала «предложение_языка», – это ошибочно построенное предложение. Всю суть процесса верификации предложения можно свести к следующему:

ЕСЛИ введенное пользователем предложение можно разложить в существующую систему правил и при этом все атрибуты в каждом правиле вывода согласуются между собой, **ТО** предложение введено пользователем верно, **ИНАЧЕ** предложение введено пользователем неверно.

Рис. 4. Метаправило, используемое разработанной системой при верификации грамматики

На рис. 5 представлен пример разложения предложения марийского языка на правила. Одной чертой подчеркнуто подлежащее, двумя чертами – сказуемое. Под каждым словом в первой строке указана часть речи, все остальные строки, расположенные ниже и обозначенные курсивом, – атрибуты.

Одна	тихая	роща	есть	в нашей	стране.
Ик	тымьк	<u>ото</u>	<u>уло</u>	мемнан	элыште.
Числительное	Прилагательное	Существительное	Глагол	Местоимение	Существительное
послед.=простое	послед.=простое	послед.=падежная_форма	послед.=изъяв_накл	послед.=простое	послед.=падежная_форма
число=ед.		падеж=именительный	время=н-б	тип=личное	падеж=местный
		число=ед.	число=ед.	падеж=родительный	число=ед.
			лицо=3-е	число=мн.	
				лицо=1-ое	

Рис. 5. Пример предложения на марийском языке

Данное предложение можно разложить в следующие правила:

$$< \text{Определение} > ::= \text{Числительное} + \text{Прилагательное} \quad (1)$$

$$< \text{Именная_группа} > ::= < \text{Определение} > + \frac{\text{Существительное}}{\text{падеж} = \text{именительный}} \quad (2)$$

$$< \text{Обстоятельство} > ::= \text{Местоимение} + \frac{\text{Существительное}}{\text{падеж} = \text{местный}} \quad (3)$$

$$< \text{Глагольная_группа} > ::= \text{Глагол} + < \text{Обстоятельство} > \quad (4)$$

$$< \text{Предложение_языка} > ::= < \text{Именная_группа} > + < \text{Глагольная_группа} > \quad (5)$$

Таким образом, если заданы правила (1)-(5), то данное предложение будет расценено системой как написанное верно. Пример верификации предложения на марийском языке показывает, что для каждого верно написанного предложения можно получить дерево вывода (см. рис. 6).

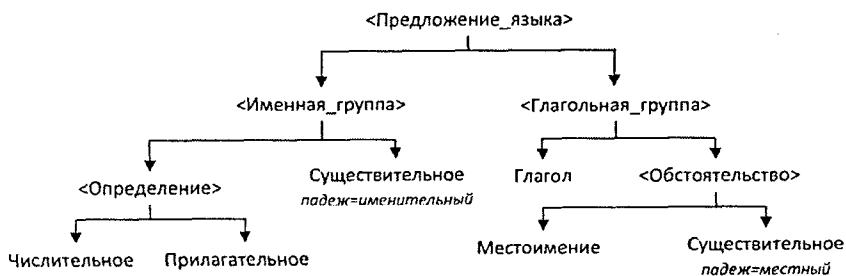


Рис. 6. Дерево вывода предложения

Предложенная модель распознавания предложения имеет свои особенности:

- наполнением правил грамматики занимаются лингвисты. Правила вводятся в систему для каждого языка отдельно через веб-интерфейс;
- исходное предложение разбивается на токены. Каждый токен получает свой набор атрибутов. Правила грамматики строятся из этих токенов с атрибутами. Атрибуты определяют семантику токена. Разбор предложения языка сводится к сведению предложения с токенами с использованием правил грамматики, введенных лингвистами, к конечному нетерминалу – символу «предложение_языка».

В третьей главе исследованы алгоритмы стемминга: алгоритм Ловинс, алгоритм Портера, алгоритм, используемый в системе И. С. Ашманова («ОРФО») и алгоритм, используемый в системе «Hunspell». Работа данных алгоритмов исследована на примере марийского и русского языков.

Показано, что данные алгоритмы имеют недостатки в ограничении длины выстраиваемых цепочек. Алгоритмы Ловинс и Портера позволяют строить цепочки, состоящие только из одного аффикса. Алгоритмы, используемые в системах И. С. Ашманова и «Hunspell», позволяют строить цепочки максимум из двух аффиксов, что недостаточно для таких языков, как марийский и финский. Кроме того, обосновано, что к недостаткам существующих алгоритмов можно отнести невозможность работать с аффиксами, которые пишутся через пробел.

Предложенный в данной работе алгоритм стемминга благодаря введению дополнительных лингвистических компонент (инфлексивов и параметров VARS) и циклической обработке цепочек аффиксов позволяет строить цепочки неограниченной длины. В общем случае генерация словоформ согласно предлагаемому алгоритму представлена на рис. 7.

Также в главе предложен алгоритм генерации подсказок для вариантов исправления ошибки в подсистеме верификации орфографии. Подсказки отображаются пользователю при осуществлении верификации орфографии текста, когда слово пользователем написано неверно. Для каждого варианта исправления ошибки выдается своя текстовая подсказка. Например, для предлагаемой словоформы «коштам» подсказка приведена на рис. 8.

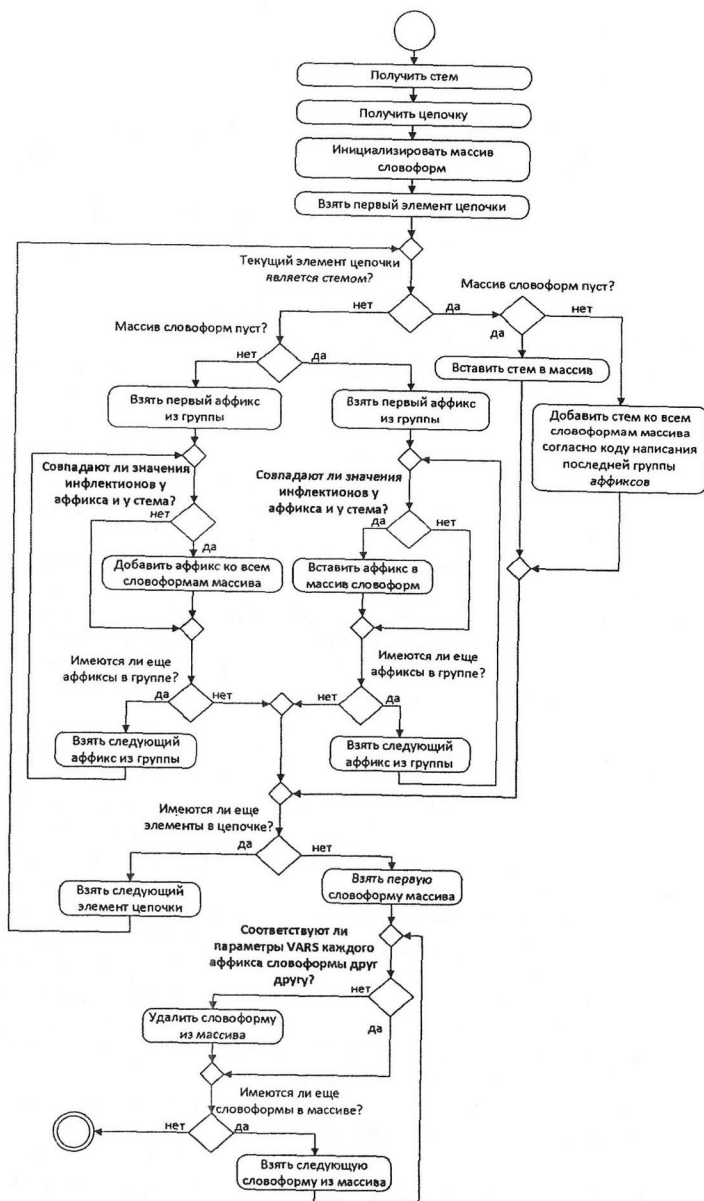


Рис. 7. Диаграмма деятельности предложенного алгоритма стемминга

коштум тыште

коштам	ге
коштым	
кошт	
кошто	

Ошибка в аффиксе. Часть речи: глагол. Изъявительное наклонение.
Время: настоящее-будущее. Лицо: первое. Число: единственное.

Рис. 8. Текстовая подсказка для предлагаемой словоформы «коштам» - одного из вариантов исправления ошибки

Текст подсказки состоит из четырех частей:

- Место совершения ошибки. На рис. 8 это «Ошибка в аффиксе».
- Часть речи. На рис. 8 это «Часть речи: глагол».
- Форма предлагаемой словоформы («Изъявительное наклонение»).
- Лингвистические характеристики предлагаемой словоформы («Время: настоящее-будущее. Лицо: первое. Число: единственное»).

Первая часть подсказки получается из сравнения последовательности аффиксов введенного пользователем слова и предлагаемого варианта исправления и может иметь три значения: ошибка в выборе формы слова, ошибка в стеме, ошибка в аффиксе.

Информация о части речи получается из цепочки аффиксов, по которой была сгенерирована словоформа. Цепочки аффиксов для каждой части речи задаются отдельно. Информация о форме предлагаемой словоформы получается из условного обозначения цепочки аффиксов. В системе хранится расшифровка условного обозначения для всех языков пользовательского интерфейса. В зависимости от выбранного пользователем языка интерфейса выбирается нужная расшифровка.

Последняя часть подсказки получается из параметров VARS. Каждый параметр имеет флаг отображения пользователю. Если этот флаг установлен в 1, то значение параметра VARS отображается пользователю в подсказке; если флаг установлен в 0, то значение параметра VARS не отображается. Также, как и с формой словоформы, в системе хранится расшифровка условных обозначений параметров и их значений, по которым и получается текстовая подсказка.

Далее в главе приводится анализ использования кодировки символов переменной длины в алгоритме Дамерау-Левенштейна. Тексты не на всех языках можно представить однобайтовыми кодировками. Среди языков, для которых отсутствуют однобайтовые кодировки, можно выделить восточные финно-угорские языки – марийский, удмуртский и коми. Тексты на этих языках можно представить кодировками символов переменной длины, в частности кодировкой UTF-8 международного стандарта Юникод.

Кодировки символов переменной длины отличаются тем, что размер памяти, занимаемый символом, варьируется от 1 до 6 байт в зависимости от самого символа. Соответственно, для корректной работы с кодировками переменной длины в автоматизированных системах необходимо каждый раз вычислять фактическую длину текущего символа, с которым производится работа. А зная всю

последовательность байт, занимаемую строкой, невозможно сразу определить длину строки в символах, не вычислив длину каждого символа в отдельности.

Предлагается модификация алгоритма Дамерау-Левенштейна, которая, благодаря вычислению длины каждого символа, позволяет вычислять меру разницы двух строк, представленных кодировкой с переменной длиной символов.

В **четвертой главе** проведен анализ программного, лингвистического обеспечения и архитектуры системы (см. рис. 3). В архитектуре выделены четыре подсистемы: подсистема взаимодействия с пользователем, подсистема верификации орфографии, подсистема верификации грамматики и подсистема наполнения лингвистических компонент. Процесс верификации орфографии и грамматики текстов обладает естественной параллельностью, так как большие объемы текста можно разбить на небольшие куски и осуществлять их параллельную верификацию. Архитектура автоматизированной системы верификации орфографии и грамматики текста отличается от существующих архитектур тем, что благодаря выделению отдельных подсистем верификации орфографии и верификации грамматики позволяет осуществлять распараллеливание процесса верификации текста большого размера.

Функциональные возможности системы в целом представлены на диаграмме вариантов использования (см. рис. 9).

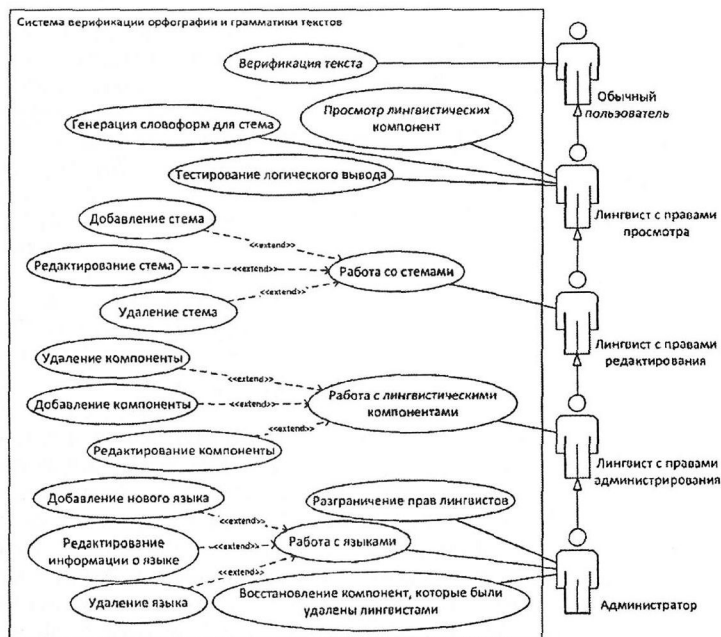


Рис. 9. Диаграмма вариантов использования разработанной системы

Далее приведена ER-диаграмма разработанной системы (рис. 10), на которой отображаются особенности хранения лингвистических компонент в базе данных системы.

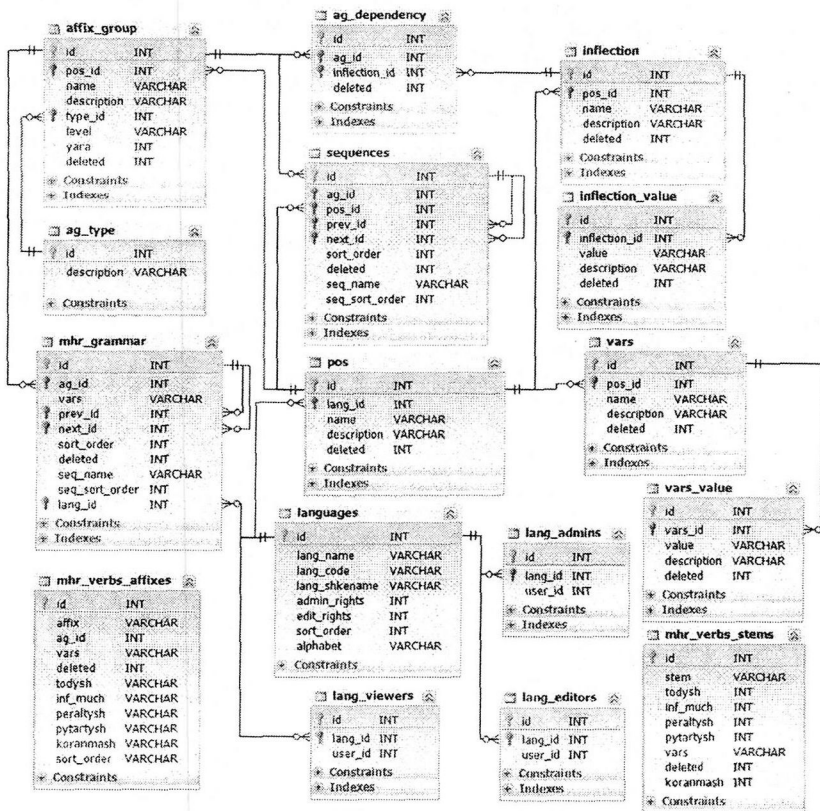


Рис. 10. ER-диаграмма разработанной системы

Обосновано использование программных интерфейсов между подсистемами системы. Информационный обмен между подсистемой взаимодействия с пользователем и подсистемами верификации орфографии и грамматики осуществляется через UNIX-сокеты, при этом все процессы выполняются на одном узле сети. Сами подсистемы верификации орфографии и грамматики работают как отдельные серверные приложения по технологии клиент-сервер. Каждый входящий запрос обрабатывается в отдельном потоке, что позволяет этим подсистемам обрабатывать несколько запросов параллельно. Информационный обмен между клиентской и серверной частью подсистемы взаимодействия с пользователем осуществляется асинхронными запросами по технологии AJAX.

Результаты работы системы показывают, что ее производительность сравнима с производительностью существующих систем. Производительность разработанной системы при работе с текстом на марийском языке составляет в среднем 46 мс для верификации одного слова на компьютере с процессором Intel Core i3 3.07GHz, 4 ГБ ОЗУ, из них 34 мс занимает информационный обмен и 12 мс – непосредственно сам процесс верификации. Верификация одного слова английского языка на компьютере с теми же самыми характеристиками системой «Hunspell» занимает 33 мс. Разработанная система для марийского языка в состоянии распознать 70 словоформ глаголов, 269-311 словоформ существительных, 2-5 словоформ прилагательных у каждого слова. В главе также приводится статистика использования системы.

В заключении сформулированы основные преимущества разработанных алгоритмов, лингвистического и программного обеспечения для систем верификации орфографии и грамматики текстов финно-угорских языков.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

- Проведен анализ лингвистических компонент языков финно-угорской группы.
- Проведены анализ и классификация методов автоматизированной верификации орфографии и грамматики текстов на языках, имеющих и не имеющих лингвистические корпуса.
- Предложены две новые лингвистические компоненты.
- Разработан алгоритм стемминга для работы с неограниченно длинными последовательностями аффиксов и для работы с частицами наравне с аффиксами.
- Предложена модификация алгоритма Дамерау-Левенштейна для вычисления меры разницы двух строк, закодированных кодировкой с переменной длиной символов.
- Разработана модель распознавания предложения исходного языка по правилам грамматики, вводимым лингвистами, для естественных языков.
- Разработан и исследован алгоритм генерации текстовых подсказок для правильных вариантов написания слова естественного языка в автоматизированных системах верификации орфографии и грамматики текстов.
- Предложены архитектура и структура программного обеспечения системы верификации орфографии и грамматики текстов финно-угорской группы.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, рекомендованных ВАК РФ

1. Килеев, В.В. Лингвистические особенности архитектуры компьютерной системы верификации орфографии финно-угорских языков / В.В. Килеев, И.Г. Сидоркина // Вестник Волжского университета имени В.Н. Татищева: науч.-теор. журнал. Серия «Информатика». – Тольятти: Волжский университет им. В.Н. Татищева, 2011. – Вып. 18. – С. 115-119.
2. Килеев, В.В. Анализ лингвистических конструкций формальной модели языка для верификации грамматики финно-угорского текста / В.В. Килеев,

И.Г. Сидоркина // Известия Кабардино-Балкарского научного центра РАН. Вып. 1 (2013) – Нальчик: Кабардино-Балкарский научный центр РАН, 2013. – С. 14-20.

3. Килеев, В.В. Кодировка символов переменной длины в алгоритме Дамерау-Левенштейна / В.В. Килеев, И.Г. Сидоркина // Вестник Чувашского университета. – Чебоксары: ЧГУ им. И.Н. Ульянова, 2013. – №3. – С. 285-292.

Публикации в иных изданиях

4. Килеев, В. В. Методы проверки орфографии марийского языка в компьютерных системах / В. В. Килеев // Информационные технологии в профессиональной деятельности и научной работе: сборник материалов Всерос. науч.-практ. конф.: в 2 ч. Ч. 2. – Йошкар-Ола: МарГТУ, 2011. – С. 35-39.

5. Килеев, В. В. Анализ алгоритмов стемминга для формализации компонентов языка финно-угорской группы / В. В. Килеев, И. Г. Сидоркина // Труды конгресса по интеллектуальным системам и информационным технологиям «IS&IT'11»: в 4 т. – М.: Физматлит, 2011. – Т.3. – С. 47-52.

6. Kileev, V. V. Models of Finno-Ugric languages' components in stemming algorithms / V. V. Kileev, I. G. Sidorkina // Interactive Systems and Technologies: the Problems of Human-Computer Interaction. Collection of scientific papers. – Ulyanovsk: UISTU, 2011. – P. 208-211.

7. Килеев, В. В. Практика использования инструментария разработки лингвистических компонентов системы проверки орфографии финно-угорских текстов / В. В. Килеев // Инновационные ресурсы и национальная безопасность в эпоху глобальных трансформаций. Пятнадцатые Вавиловские чтения; постоянно действ. Всерос. междисциплинар. науч. конф. с междунар. участием: в 2 ч. / редкол.: В. П. Шаласов и др. – Йошкар-Ола: МарГТУ, 2012. – Ч. 2. – С. 267-268.

8. Килеев, В. В. Схема функционирования системы верификации грамматики языков / В. В. Килеев // Информационные технологии в профессиональной деятельности и научной работе: сб. материалов Всерос. науч.-практ. конф. с междунар. участием: в 2 ч. – Йошкар-Ола: МарГТУ, 2012. – Ч.2. – С. 107-111.

9. Килеев, В. В. Моделирование синтаксических и грамматических правил для системы верификации текста / В. В. Килеев // Информатика и вычислительная техника: сб. науч. трудов 4-й Всерос. науч.-техн. конф. аспирантов, студентов и молодых ученых: в 2 т. Т. 1 / под ред. Н. Н. Войта. – Ульяновск: УлГТУ, 2012. – С. 302-304.

10. Килеев, В. В. Модель представления правил для системы верификации грамматики текстов языков финно-угорской группы / В. В. Килеев, И. Г. Сидоркина // Материалы третьей междунар. конф. «Автоматизация управления и интеллектуальные системы и среды». 9-15 октября, Махачкала, Россия. Т.2. – Нальчик: Издательство КБНЦ РАН, 2012. – С. 159-163.

11. Килеев, В. В. Компоненты архитектуры компьютерной системы верификации орфографии финно-угорских языков / В. В. Килеев // Программные системы и вычислительные методы. – М.: Nota bene, 2012. – № 1. – С. 37-42.

12. Килеев, В. В. Методы верификации грамматики естественных языков финно-угорской группы на уровне семантического представления / В. В. Килеев, И. Г. Сидоркина // Открытые семантические технологии проектирования

интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2013): материалы III междунар. науч.-техн. конф. (Минск, 21-23 февраля 2013 года) / редкол.: В.В. Голенков (отв. ред.) [и др.]. – Минск: БГУИР, 2013. – С. 333-336.

13. Килеев, В. В. Модификация алгоритма Дамерау-Левенштейна для работы с символами переменной длины нечетких моделей компьютерной лингвистики / В. В. Килеев, И. Г. Сидоркина // Интегрированные модели и мягкие вычисления в искусственном интеллекте: сб. науч. трудов VII-й Междунар. науч.-техн. конф. (Коломна, 20-22 мая 2013 г.): в 3 т. Т.3. – М.: Физматлит, 2013. – С. 1249-1256.

14. Килеев, В. В. Лингвистические компоненты языка в системе верификации орфографии и грамматики, не использующей лингвистические корпуса / В. В. Килеев // Информационные технологии в профессиональной деятельности и научной работе: сб. материалов Всерос. науч.-практ. конф. с междунар. участием: в 2 ч. Ч. 2. – Йошкар-Ола: ПГТУ, 2013. – С. 125-129.

15. Килеев, В. В. Новые лингвистические компоненты алгоритма стемминга в системе верификации орфографии финно-угорских языков / В. В. Килеев, И. Г. Сидоркина // Труды Конгресса по интеллектуальным системам и информационным технологиям «IS&IT'13»: в 4 т. – М.: Физматлит, 2013. – Т. 1. – С. 140-147.

Подписано в печать 18.10.2013. Формат 60×84/16.

Бумага офсетная. Печать офсетная. Усл. печ. л. 1,0. Тираж 100 экз. Заказ № 5209.
Редакционно-издательский центр ПГТУ. 424006 Йошкар-Ола, ул. Панфилова, 17

10 —