Желтов Валерьян Павлович,

к.т.н., профессор;

Желтов Павел Валерианович,

к.т.н., доцент;

Губанов Алексей Рафаилович,

д.ф.н., профессор;

Пушкин Александр Сергеевич,

магистрант 2 курса ИВТ,

ФГБОУ «Чувашский государственный университет им. И.Н. Ульянова»,

г. Чебоксары, Чувашская Республика

ПОДСИСТЕМА АНАЛИЗА ТЕКСТОВ В ПОИСКОВИКЕ ДЛЯ НАЦИОНАЛЬНОГО КОРПУСА ЧУВАШСКОГО ЯЗЫКА

Аннотация. В статье рассмотрена подсистема анализа текстов в поисковике. На данном этапе подсистема анализа текстов состоит из следующих компонент: 1) компоненты токенизации текста; 2) компонента выделения предложений в тексте; 3) компоненты морфологического анализа предложений. Для хранения лингвистических данных, полученных в результате работы компонент поисковика, необходимы следующие специальные структуры данных в виде набора классов, описанная в статье. Компонента токенизации текста преобразует текст в набор токенов (слов, сокращений и т.д.). Для задания правил токенизации используется файл настройки, содержащий регулярные выражения и список слов сокращений. Для выделения предложений в тексте используется файл настройки, в котором указывается необходимость разбиения на предложения текста с помощью ряда правил, примеры которых приведены в статье.

Ключевые слова: поисковик, текстовый корпус, разметка текста, запрос, индексирование.

В настоящее время в компьютерной лингвистике одной из актуальных задач является создание электронных национальных корпусов. Подобные корпуса, представляют собой электронные библиотеки размеченных текстов с возможностью быстрого поиска на нескольких уровнях языка: морфемном, морфологическом, синтаксическом, текстовом и семантическом. Подобные текстовые корпуса созданы уже для многих языков Российской Федерации (русский, татарский, башкирский, калмыцкий, марийский, мордовский,

удмуртский, коми, хакасский). В настоящее время авторами данной статьи ведется работа над созданием национального корпуса чувашского языка.

Подобные корпуса обслуживаются большим количеством программных продуктов, которые позволяют производить их обработку и выполняют различные пользовательские запросы, направленные на исследования текстов и выборку из них определенных интересующих пользователя данных.

Одним из основных программных продуктов в национальном корпусе является поисковик. В поисковик, в свою очередь, может входить множество модулей, одним из которых является подсистема анализа текстов.

Рассмотрим подсистему анализа текстов в поисковике. На данном этапе подсистема анализа текстов состоит из следующих компонент**:** 1) компоненты токенизации текста**;** 2) компонента выделения предложений в тексте**;** 3) компоненты морфологического анализа предложений.

Для хранения лингвистических данных, полученных в результате работы компонент поисковика, необходимы следующие специальные структуры данных (специальные классы):

• Word (слово) – словоформа со списком возможных объектов – результатов анализа (Analysis). Поля класса Word представлены в таблице 1.

Таблица 1

InDict	Флаг, указывающий на то, что словоформа была найдена в словаре
Form	Словоформа
Start	Смещение начала токена в исходном сообщении
Finish	Смещение конца токена в исходном сообщении
User	Пользовательские данные
Analyses	Список объектов - результатов анализа

• Analysis (анализ) – с каждым словом соотносятся результаты морфологического анализа (может быть несколько альтернативных вариантов результатов в силу неопределенности и двусмысленности, присутствующей в тексте), подробнее см. в [1], [2]. Поля класса Analysis представлены в таблице 2.

Таблица 2

Lemma	Гнездовое слово
Tag	PoS тег (Part-of-speech)/ тег части речи
Descriptor	Дескриптор
Probability	Вероятность того, что у словоформы (объекта Word) действительно такая лемма/тег.
User	Пользовательские данные

• Sentence (предложение) – класс содержит список слов, составляющих законченное предложение.

• Document (документ) – класс содержит список предложений, составляющих сообщение эксперта.

Отметим, что компонента токенизации текста преобразует текст в набор токенов (слов, сокращений и т.д.). Для задания правил преобразования используется файл настройки, содержащий регулярные выражения и список слов сокращений.

Правило токенизации состоит из двух частей: имени правила и регулярного выражения, которое используется для выделения токена. Примеры правил токенизации показаны в таблице 3:

Таблица 3

Регулярное выражение	Имя правила
WORD {[[:alnum:]°°]}+[\+]*	Правило для выделения слов
TIMES (([01]?[0-9]|2[0-4]):[0-5][0-9])	Правило для выделения из текста времени

Примеры сокращений (трактуемых как один токен) ниже в таблице 4:

Таблица 4

Сокращение	Значение
арифм.	арифметическое
д.ф.-м.н.	доктор физико-математических наук

Компонента выделения предложений получает на вход список токенов и возвращает список предложений.

Для выделения предложений используется файл настройки, в котором: указывается необходимость разбиения на предложения текста, который находится между парой маркеров (открывающим маркером и закрывающим маркером); приводится список пар «открывающе-закрывающих» маркеров (это пары символов (или пары групп символов) такие как «[« и »]», «{« и »}» и т.д.); приводится список возможных начальных и конечных символов предложения (например, «.», «!», «?»), причем для конечных символов указывается необходимость анализа последующего заглавного символа или начального символа.

Модуль поиска по словарю ищет заданное слово и возвращает леммы и соответствующие им части речи. Файл словаря форм представляет собой обычный текстовый файл, состоящий из текстовых строк. В текстовых строках содержатся формы слов в формате «форма лемма1 часть речи1 | лемма2 часть речи2 | ...».

Сокращения, соответствующие частям речи, основаны на имеющемся наборе тегов разметки национального корпуса чувашского языка, частично описанного в [4].

Словарь был создан на базе инверсионного, грамматического словаря – Обратного словаря чувашского языка [5], ибо практическая значимость словарей такого типа заключается в группировке слов по одинаковому концу: для чувашского языка данный принцип особенно важен, так как аффиксы в нем располагаются справа от корня. Слова в инверсионном словаре в дальнейшем можно сгруппировать по морфологическому признаку (часть речи, наличие или отсутствие того или иного аффикса). В частности, анализ существующих обратных словарей на практике позволил нам представить многообразие аффиксальных средств имен в чувашском языке, их продуктивность. В обратном словаре имеются массивы слов (более тысячи), которые имеют определенный аффикс.

Обратный словарь с программной точки зрения может функционировать и отдельно как база данных и диалоговый интерпретатор запросов к базе

данных в разработанной нами системе поисковика. Обратный словарь с перечисленными нами характеристиками, т.е. с наиболее полной информацией о грамматической характеристике лексики чувашского языка может быть использован для квантитативного описания по широкому кругу морфологических характеристик.

Модуль поиска по словарю определяет при помощи данного словаря априорную вероятность каждого возможного анализа каждой словоформы в предложении (естественно, только в случае наличия нескольких вариантов анализа). Если же для слова не был определен результат анализа, то модуль пытается угадать возможные теги PoS (части речи) слова, основываясь на окончании слова.

Модули представляют собой конечные автоматы, используемые для выделения чисел и дат в тексте.

На основании указанных модулей была также создана отдельная информационная система. «Лексический поисковик» предназначена для поиска и анализа предложений художественных произведений, содержащих указанные пользователем ключевые слова. Разрабатываемая нами система состоит из следующих компонентов или модулей:

- Модуль управления пользовательским интерфейсом. Модуль принимает запросы пользователя, направляет запросы другим модулям и выдает результаты выполнения запросов пользователю.
- Модуль индексирования и поиска текстов. Модуль на основании выбранных пользователем ключевых слов находит все релевантные предложения из базы индексов художественных текстов, затем выдает их пользователю, используя при этом базу структурированных данных (в дополнение к предложению пользователю выдаются автор художественного текста, название художественного текста и т.д.).
- Модуль анализа текстов используется всеми остальными модулями. Модуль позволяет проводить лексический, морфологический и синтаксический анализ текстов [3].

При запуске системы загружается стартовая форма, которая состоит из нескольких областей. Левая область «Поиск» состоит из текстовых полей ввода: «Автор произведения» (предложения будут найдены из произведений только указанных авторов), «Название произведения» (предложения будут найдены только из указанных произведений), «Ключевые слова» (будут найдены предложения, содержащие указанные ключевые слова). Пользователь может в каждом поле использовать логические связки И/ИЛИ/НЕ, вложенные скобки, а также специальные мета-символы * (замещает любое количество букв) и ? (замещает одну букву). Заполнив требуемые поля, пользователь может нажать на кнопку «Найти» (находится в той же левой панели); при этом запустится модуль индексирования и поиска текстов (и опосредованно модуль анализа текстов). Релевантные пользовательскому запросу предложения (с дополнительной метаинформацией) отобразятся в правой верхней области «Художественные произведения».

Далее пользователь может выбрать интересующее его художественное произведение (можно использовать сортировку по полям метаинформации: автору произведения, названию произведения, дате публикации и т.д.) и дважды щелкнуть на нем, после чего в правую среднюю область «Художественное произведение» загрузятся все искомые предложения (точнее список предложений, содержащих указанные пользователем ключевые слова), для каждого предложения можно посмотреть содержащий его текстовый контекст.

На созданной интегрированной базе планируется разработка системы «Автоматическое рабочее место (АРМ) лексикографа», которая в своем составе имела бы не только электронную картотеку, но и текстотеку и лингвостатистический пакет, для статистического анализа лексики и Национального корпуса текстов на основе статистических методов (корреляционного, дисперсионного, факторного и др.).

В разрабатываемом поисковике определены и внедрены все пакеты системы и основные классы пакетов.

В пакете лексического поисковика разработаны классы каркаса пользовательского приложения, ответственного за обработку событий и взаимодействие с пользователем (Рис. 1).

Выводы

Одной из основных проблем усложняющих работу поисковика является нестандартизованность чувашской орфографии, а именно непрекращающиеся споры по вопросу слитного или раздельного написания чувашских аналитических, в том числе изафетных конструкций, из которых в чувашском, как и в других тюркских языках, строятся новые понятия.

Указанные особенности были учтены в созданной лексикографической базе инверсионного, грамматического словаря. Для этого аналитические конструкции даются в двух формах: слитной и раздельной.

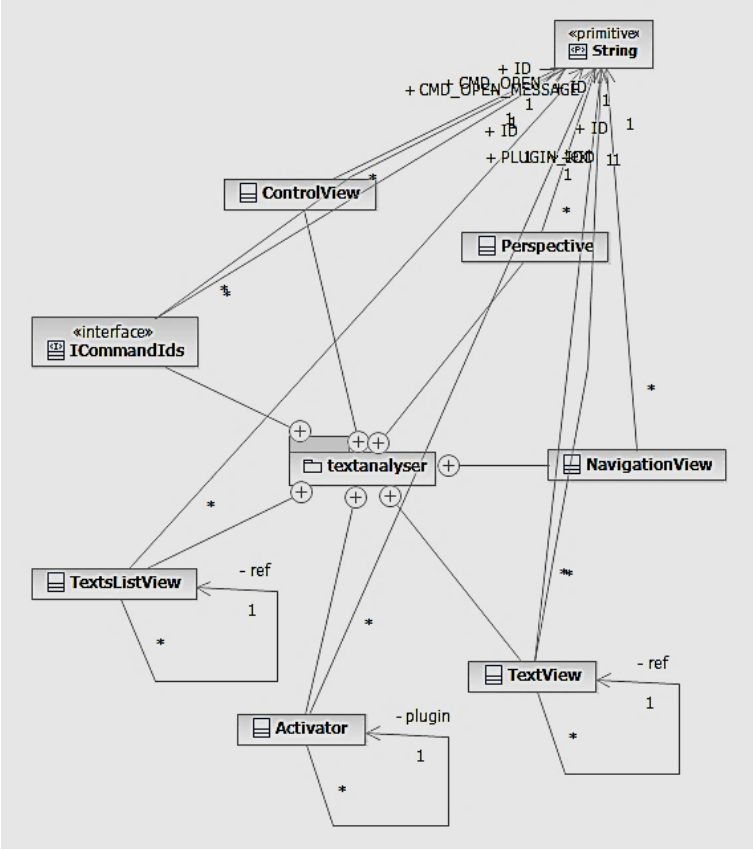

Рисунок 1 – Взаимосвязи между классами пакета

В ходе работы над поисковиком нами были выявлены также некоторые проблемы грамматической классификации, отражающие особенности тюркских

языков вообще и чувашского языка в частности. Так, нечеткость границ между словоизменительными классами в ходе разработки алгоритмов морфологической разметки чувствуется не только в пределах имени, но и в пределах других частей речи. Например, если к имени в чувашском языке присоединяются показатели категории выделения, то оно начинает выполнять предикативную функцию; в предложении прилагательное в роли актанта может принимать именные показатели.

Также при автоматической разметке возникают трудности, связанные с полифункциональными и омонимичными аффиксами, с особенностями изменения отдельных категорий слов (в частности, некоторых местоимений и послеложных слов), с частеречной принадлежностью слов в безаффиксальной форме, с числом падежей и т.д.

Лексические формы, относящиеся традиционно к разным морфологическим классам, часто принимают аффиксы одних и тех же морфологических категорий, что для программы является формальным основанием отнести их к одному словоизменительному классу.

В целом указанные проблемы можно решить за счет разработки и применения грамматики порядков, определяющей фиксированную последовательность словообразовательных аффиксов и однократное появление в той или иной словоформе аффикса определенной граммемы. Чувашский язык имеет, что важно для грамматической разметки, развитую систему грамматически однозначных словоизменительных аффиксов (отдельно взятый аффикс выражает, как правило, один морфологический признак, в нем отсутствуют различные парадигматические классы в парадигме того или иного одного типа; имеется также закономерная фонетическая обусловленность алломорфов (границы морфем чёткие: к основе присоединяются аффиксы с тем или иным значением, а если происходят фонемные изменения на границах морфем, то данные морфонологические изменения связаны с фонологическими законами чувашского языка). При автоматическом анализе аффиксального состава словоформ в рассматриваемом языке грамматические

(морфологические) признаки распознаются относительно легко. В общем и целом, чувашская морфология вписывается в общую схему категорий и форм присущих тюркским языкам, а чувашский аналитические конструкции являются типично тюркскими.

СПИСОК ЛИТЕРАТУРЫ

1. Желтов П.В. Лингвистические процессоры, формальные модели и методы: теория и практика / П.В. Желтов. – Чебоксары: Изд-во Чуваш. ун-та, 2006. – 208 с.

2. Желтов П.В. Формальные методы в сравнительно-сопоставительном языкознании / П.В. Желтов. – Чебоксары: Изд-во Чуваш. ун-та, 2006. – 252 с.

3. Желтов П.В. Лингвистические процессоры в системах искусственного интеллекта / П.В. Желтов. – Чебоксары: Изд-во Чуваш. ун-та, 2007. – 100 с.

4. Zheltov Pavel. Morphological markup system for the national body of the Chuvash language / Pavel Zheltov // Proceedings of the International conference «Turkic Languages Processing: TurkLang 2015». – Kazan: Academy of Sciences of the Republic of Tatarstan Press, 2015. – pp. 328-330.

5. Zheltov Pavel. Reverse Dictionary of Chuvash. Обратный словарь чувашского языка / Pavel Zheltov, Eduard Fomin, Jorma Luutonen // Société Finno-Ougrienne. – Helsinki. 2009. – 344 p.

Valeryan P. Zheltov,

Candidate of engineering sciences, professor;

Pavel V. Zheltov,

Candidate of engineering sciences, associate professor;

Aleksey R. Gubanov,

Doctor of philological sciences, professor;

Aleksandr S. Pushkin,

2'nd year Master's Degree Student at The Faculty of Information-computer Technologies,

FSEBI of HE «The Ulianov Chuvash State University»,

Cheboksary, Chuvash Republic

THE SUBSYSTEM OF TEXT ANALYSIS IN THE SEARCH ENGINE FOR THE NATIONAL CORPORA OF THE CHUVASH LANGUAGE

Abstract. The article considers the subsystem of text analysis in the search engine. At this stage, the text analysis subsystem consists of the following components: 1) text tokenization

components; 2) the component of the selection of sentences in the text; 3) the components of the morphological analysis of sentences. To store the linguistic data obtained as a result of the search engine component, the following special data structures are required in the form of a set of classes, described in the article. The text tokenization component converts text into a set of tokens (words, abbreviations, etc.). To set the rules of tokenization used a configuration file that contains regular expressions and a list of abbreviations. To highlight sentences in the text, a configuration file is used, which indicates the need to break up the sentence of the text using a series of rules, examples of which are given in the article.

Keywords: search engine, text corpora, query, indexing.