

Министерство образования и науки Российской Федерации  
Федеральное государственное автономное образовательное учреждение  
высшего профессионального образования  
«Уральский федеральный университет имени первого Президента России Б. Н. Ельцина»

Физико–технологический институт

Кафедра вычислительной техники

ДОПУСТИТЬ К ЗАЩИТЕ В ГЭК

Зав. кафедрой, д. т. н., профессор

\_\_\_\_\_ С. Л. Гольдштейн

«\_\_\_\_\_» \_\_\_\_\_ 2011 г.

**ИССЛЕДОВАНИЕ И РАЗРАБОТКА СИСТЕМЫ  
АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ  
КЛЮЧЕВЫХ ФРАЗ ИЗ ТЕКСТА  
НА ЕСТЕСТВЕННОМ ЯЗЫКЕ**

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА  
БАКАЛАВРА**

**ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

Руководитель доц., к. ф.-м. н. \_\_\_\_\_ А. Г. Кудрявцев

Консультант зав. сектором, к. т. н. \_\_\_\_\_ А. В. Созыкин

Консультант доц., к. т. н. \_\_\_\_\_ П. И. Браславский

Нормоконтролёр доц., к. т. н. \_\_\_\_\_ В. В. Ковалёв

Студент гр. ФТ-47081 \_\_\_\_\_ Д. А. Усталов

Екатеринбург

2011

## Реферат

Пояснительная записка: 74 страниц; 23 рисунка; 10 таблиц; 48 источников; 2 приложения.

Разработано программное средство, извлекающее список ключевых фраз из текста на естественном языке.

В процессе выполнения работы проведён аналитический обзор существующих систем автоматического извлечения ключевых фраз из текста на естественном языке. Выбран прототип. Проведён анализ прототипа и предложены пути его развития. Разработан пакет моделей и составлено Техническое задание. Выполнено проектирование и получена инженерная реализация Tesuĉk — системы автоматического извлечения ключевых фраз из текста на естественном языке.

Программный пакет создан на языке Ruby.

# Содержание

Введение . . . . .	5
1 Проблематика систем автоматического извлечения ключевых фраз из текста на естественном языке . . . . .	6
1.1 Основные термины и понятия . . . . .	6
1.2 Технология поиска информации . . . . .	6
1.3 Обзор аналогов . . . . .	7
1.3.1 OpenCalais . . . . .	7
1.3.2 Extractor . . . . .	9
1.3.3 Yahoo! Term Extraction Web Service . . . . .	11
1.3.4 TerMine . . . . .	13
1.3.5 Maui . . . . .	15
1.3.6 TextAnalyst . . . . .	17
1.3.7 AOT . . . . .	19
1.3.8 ContentAnalyzer . . . . .	21
1.3.9 Семантическое зеркало . . . . .	23
1.4 Формирование набора критериев . . . . .	25
1.5 Сравнение аналогов . . . . .	25
1.6 Работа с прототипами . . . . .	26
1.6.1 Критика прототипа . . . . .	29
1.6.2 Предлагаемое решение . . . . .	30
1.6.2.1 Обзор морфологических анализаторов . . . . .	33
1.7 Результаты и выводы . . . . .	36
2 Модель предлагаемого решения . . . . .	37
2.1 Содержательная модель . . . . .	37
2.2 Концептуальная модель . . . . .	37
2.2.1 Общая модель . . . . .	37
2.2.2 Базово–уровневая модель . . . . .	38
2.2.3 Модификационная модель . . . . .	38
2.3 Структурная модель . . . . .	39
2.4 Функционально–структурная модель . . . . .	40
2.5 Алгоритмическая модель . . . . .	40
2.6 Результаты и выводы . . . . .	40

3	Проектирование предлагаемого решения . . . . .	42
3.1	Внешнее проектирование . . . . .	42
3.2	Внутреннее проектирование . . . . .	42
3.3	Результаты и выводы . . . . .	42
4	Инженерная реализация и эксплуатация предлагаемого решения . . . . .	43
4.1	Требования к средствам обеспечения . . . . .	43
4.2	Экранные формы . . . . .	44
4.3	Результаты и выводы . . . . .	45
	Заключение . . . . .	47
	Список использованных источников . . . . .	49
А	Функционально–структурные модели . . . . .	53
Б	Техническое задание . . . . .	57

## Введение

Задача выделения ключевых слов и фраз из текста возникает в библиотечном деле, лексикографии и терминоведении, а также в информационном поиске. В настоящее время, объёмы и динамика информации, которая подлежит обработке в этих областях, делают особенно актуальной задачу *автоматического выделения* ключевых слов и фраз. Выделенные таким образом слова и словосочетания могут использоваться для создания и развития терминологических ресурсов, а также для эффективной обработки документов: индексирования, реферирования и классификации [1].

Исследованию данной задачи посвящено множество работ [2–5], предлагающих различные методы выделения ключевых слов и фраз из текста, построенные как на основе статистических, так и лингвистических моделей.

Существует большое число доступных систем автоматического выделения ключевых фраз, разработанных и ориентированных исключительно на обработку западноевропейских языков [6–10]. Очевидно, поддержка русского языка в таких системах отсутствует.

Системы автоматического извлечения ключевых фраз из текста на естественном языке также разрабатывались в России [11–14], но из-за особенностей их программной реализации, применения устаревших и недостаточно эффективных методов извлечения ключевых фраз [1, 15, 16] и условий распространения не могут быть использованы для решения обозначенных выше прикладных задач.

Сегодня интеллектуальные информационные системы нашли широкое применение в области здравоохранения. Одной из важнейших составляющих современных медицинских информационных систем является подсистема интеллектуального анализа текстовых данных, адекватность функционирования которой напрямую зависит от качества работы модуля автоматического извлечения ключевых фраз.

Таким образом, актуальны задачи исследования и разработки *системы автоматического извлечения ключевых фраз из текста на естественном языке*.

# 1 Проблематика систем автоматического извлечения ключевых фраз из текста на естественном языке

## 1.1 Основные термины и понятия

**Ключевое слово** (англ. *keyword*) — слово в тексте, способное в совокупности с другими ключевыми словами представлять текст.

**Ключевая фраза** (англ. *keyphrase*) — выражение, состоящее из одного или нескольких ключевых слов, представляющее собой важнейший информационный сегмент документа.

В дальнейшем, под **термином** будет пониматься *ключевая фраза*.

**Графематический анализ** — этап обработки текста, на котором выполняется разбиение текста на заголовки, абзацы, предложения, обороты, отдельные слова и цифро–буквенные комплексы.

**Морфологический анализ** — этап обработки текста, на котором проводится построение морфологической интерпретации слов исходного текста и определение их грамматических характеристик, таких как число, род, падеж, и т. д.

**Синтаксический анализ** — этап обработки текста, на котором определяются синтаксические связи между словами в каждом предложении исходного текста.

## 1.2 Технология поиска информации

Поиск информации осуществлялся в трёх направлениях:

— **Интернет**: использовались поисковые системы Google и Яндекс, а также материалы из Википедии.

— **Эксперты**: были опрошены научные сотрудники отдела вычислительной техники ИММ УрО РАН; специалисты компании Яндекс в области компьютерной лингвистики и информационного поиска.

— **Печатные издания**: были изучены публикации в сборниках трудов международных конференций, посвящённых компьютерной лингвистике и искусственному интеллекту; публикации в журналах, входящих в список ВАК,

имеющие УДК 004.912 и 004.8; книги, посвящённые компьютерной лингвистике.

### 1.3 Обзор аналогов

В настоящее время существует большое количество систем автоматического извлечения ключевых фраз из текста на естественном языке.

Ключевыми факторами при отборе аналогов в данной работе были рекомендации экспертов, количество исследовательских работ, посвящённых аналогам, а также популярность соответствующих систем в современном IT-сообществе.

#### 1.3.1 OpenCalais

OpenCalais — Web-сервис, предназначенный для автоматического извлечения семантических метаданных из текстов на естественном языке [6]. Начиная с 2007 года развитием и поддержкой сервиса занимается корпорация Thomson Reuters.

Семантические метаданные представлены в виде именованных сущностей (англ. *named entity*), а также связанных с ними фактов и событий. Именованные сущности, в свою очередь, могут рассматриваться как ключевые слова и фразы исходного текста.

Пример использования системы приведён на рисунке 1.1. Подчёркиванием выделены слова, определённые системой как именованные сущности исходного текста. Цвет подчёркивания обозначает тематическую принадлежность каждого выделенного термина.

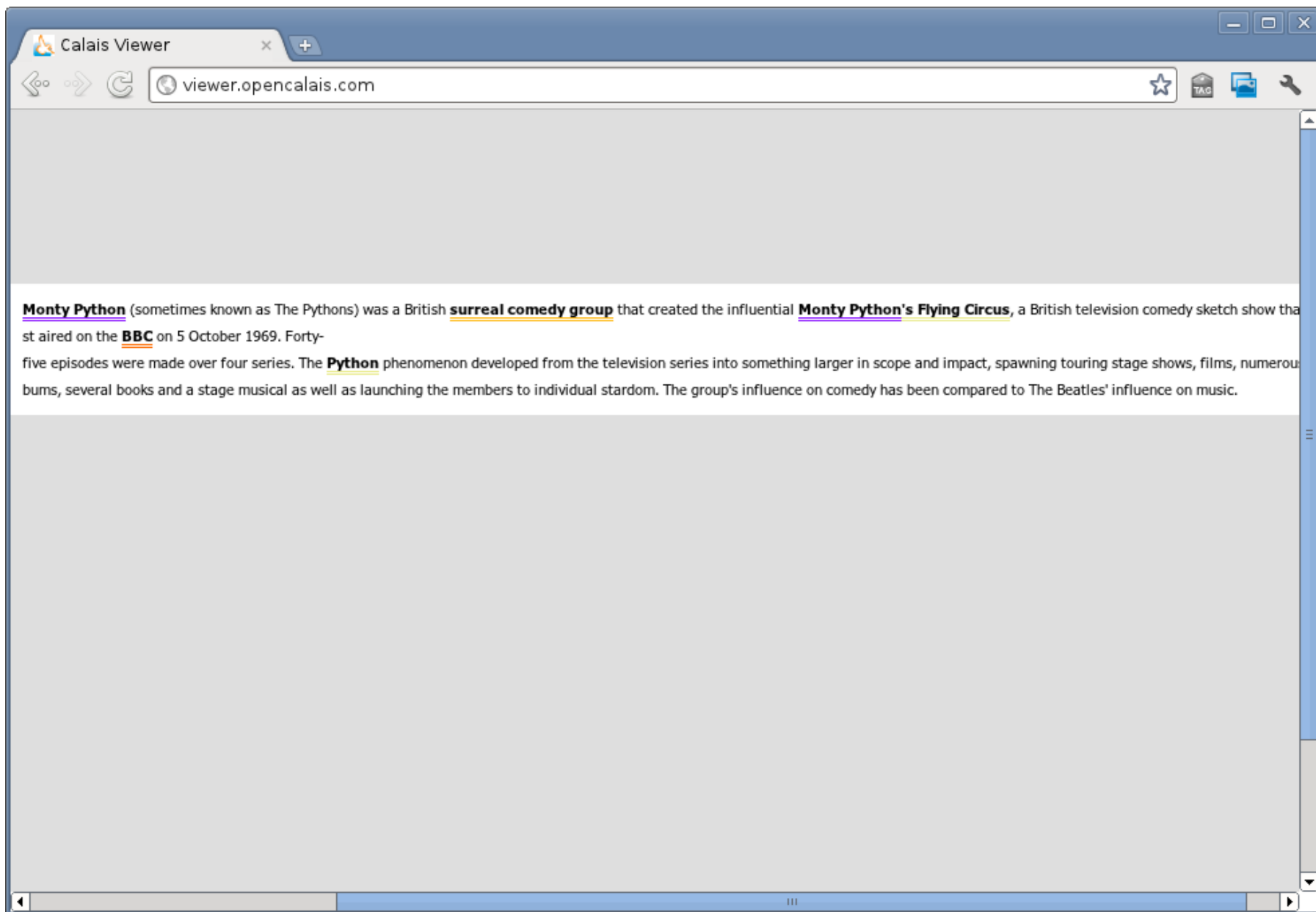


Рисунок 1.1 — Пример использования Web-приложения Calais Viewer

Функционирование системы OpenCalais основано на методах обработки естественного языка, машинного обучения и других алгоритмах.

Для извлечения семантических метаданных применяются предварительно подготовленные онтологии различных предметных областей в формате RDF. Исходный текст подвергается предварительной обработке (графематической и морфологической разметке), затем размеченные словосочетания проходят идентификацию при помощи обученной модели распознавания именованных сущностей, между которыми ведётся поиск семантических отношений. Полученный граф сущностей и отношений между ними преобразуется в набор RDF-троек.

Web-сервис OpenCalais бесплатен и доступен для некоммерческого и коммерческого использования, однако требует регистрации для получения API-ключа. Сервис построен по архитектуре REST [17] и использует формат XML [18] для обмена данными с пользователями.

На сегодняшний день, Web-сервис OpenCalais не способен обрабатывать русскоязычные тексты.

### 1.3.2 Extractor

Extractor — система автоматического извлечения терминов, функционирующая с 2002 года и используемая многими организациями в собственных решениях по обработке естественного языка [7].

Работа системы Extractor основана на применении генетических алгоритмов [19] в сочетании с методами машинного обучения и статистическими методами обработки естественного языка [3]. Первоначальное обучение системы ведётся на основе размеченного корпуса текстов.

Ознакомиться с возможностями Extractor можно при помощи демонстрационного Web-приложения [ExtractorLive.com](http://ExtractorLive.com). Для создания собственных решений на основе технологии Extractor необходимо приобрести набор инструментов разработчика. Пример использования [ExtractorLive.com](http://ExtractorLive.com) приведён на рисунке 1.2. Под заголовком “Keyphrases” выведен список терминов, извлечённых из исходного текста, а под заголовком “Highlights” показан результат автореферирования исходного текста.

На сегодняшний день, Extractor не способен обрабатывать русскоязычные тексты.

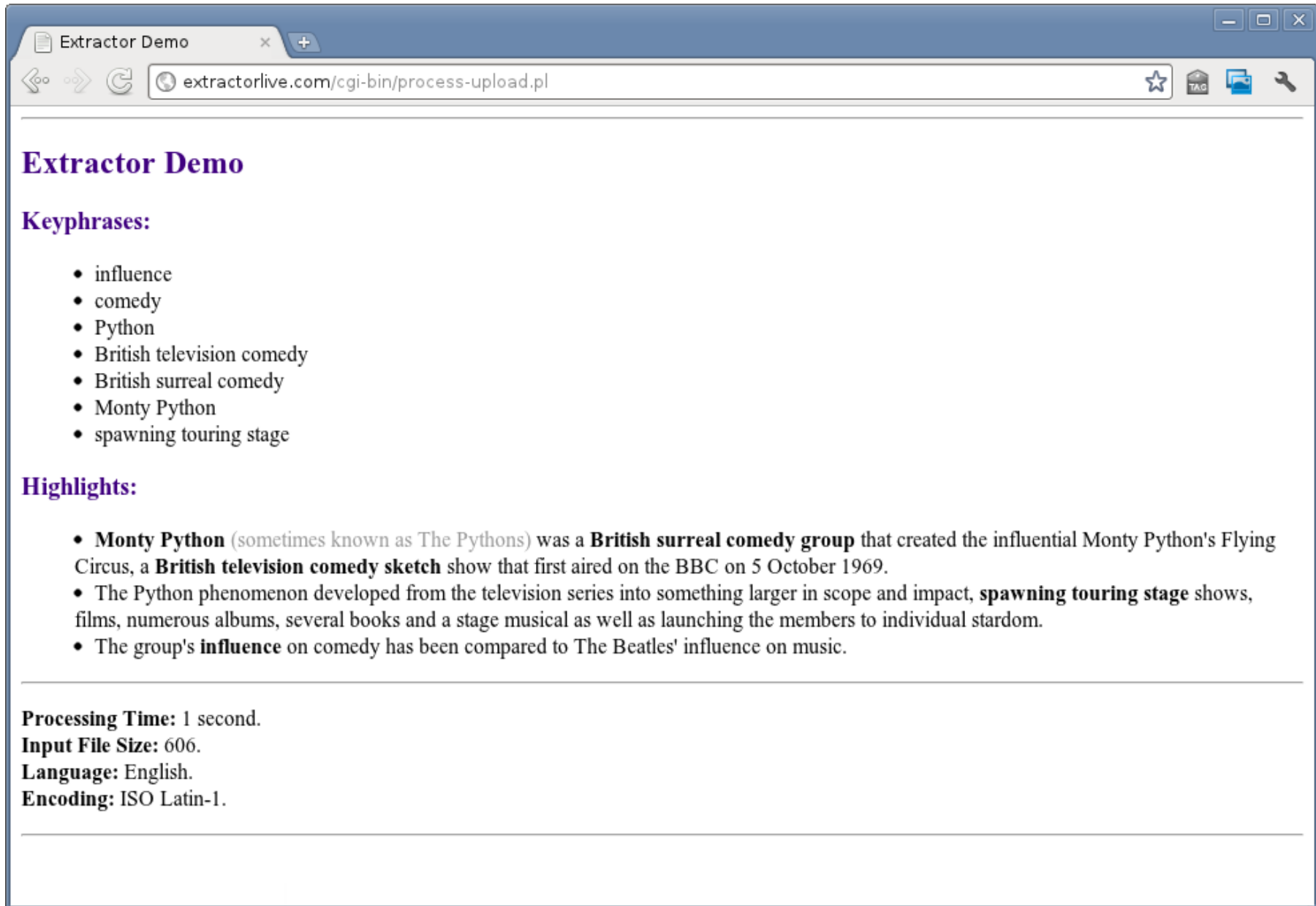


Рисунок 1.2 — Пример использования Web-приложения ExtractorLive.com

### 1.3.3 Yahoo! Term Extraction Web Service

Yahoo! Term Extraction Web Service — сервис, используемый в поисковой системе Yahoo! Search, предназначенный для извлечения ключевых фраз из текста на естественном языке [8].

Документация Yahoo! Term Extraction Web Service не упоминает используемую технологию извлечения терминов.

Программный интерфейс Yahoo! Term Extraction Web Service доступен наряду с другими сервисами в составе системы YQL. Обмен данными с пользователем осуществляется в популярных форматах XML и JSON.

В данный момент, обработка русскоязычных текстов при помощи Yahoo! Term Extraction Web Service невозможна.

Пример использования системы приведён на рисунке 1.3. Результат работы Yahoo! Term Extraction Web Service представлен в виде XML-дерева и показан на вкладке “Tree” в виде списка кандидатов в термины.



### 1.3.4 TerMine

TerMine — Web-сервис извлечения терминов, разработанный в британском Национальном центре анализа текста (англ. *The National Centre for Text Mining*) [9].

Сервис TerMine работает на основе метода C-value, описанного в [4] и применяет анализатор TreeTagger [20] для предварительной морфологической разметки текста.

Демонстрационный Web-интерфейс TerMine позволяет обрабатывать тексты объёмом до 2 мегабайт исключительно на английском языке.

Пример использования системы приведён на рисунке 1.4: термины, извлечённые из исходного текста, выделены красным цветом.

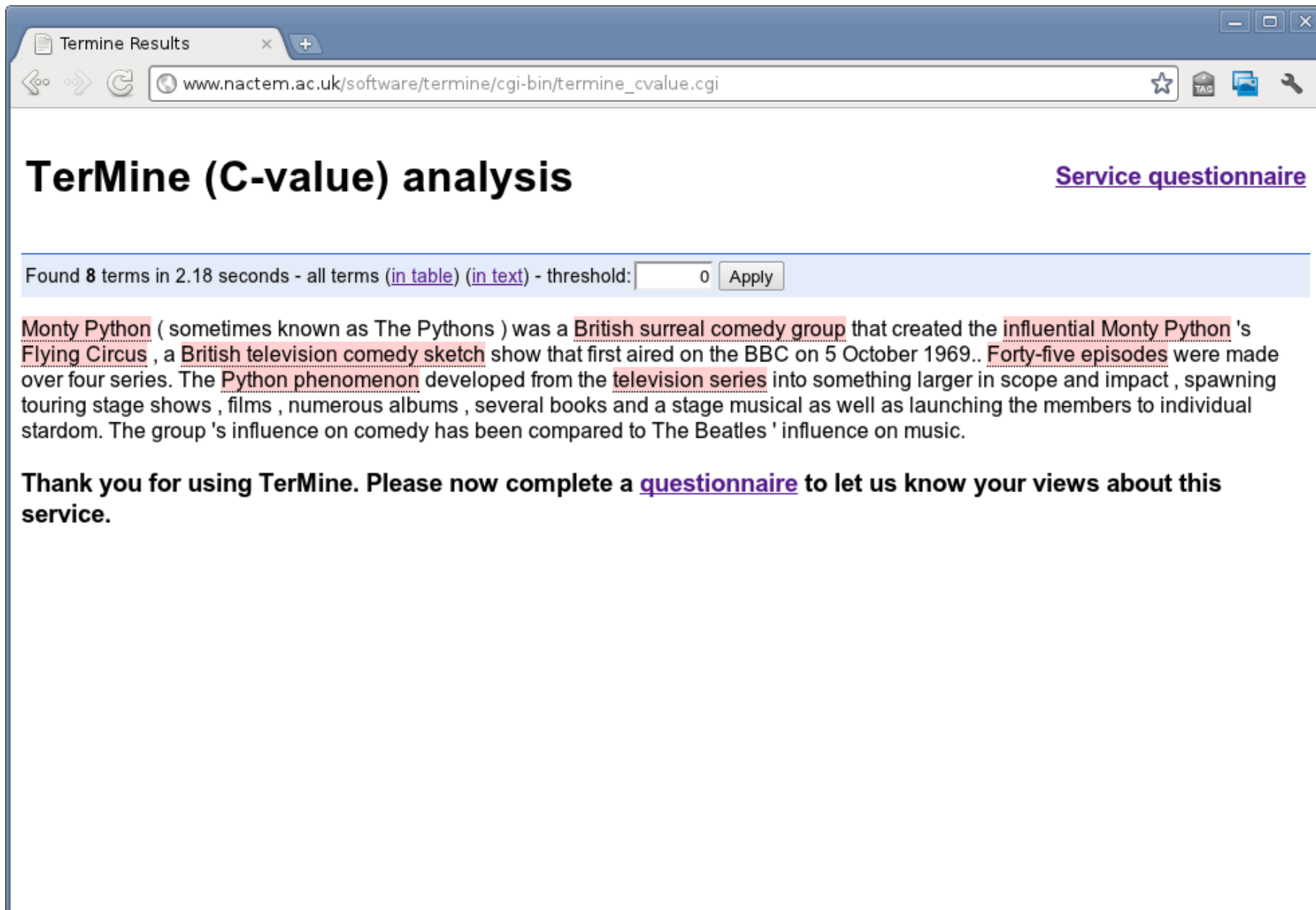


Рисунок 1.4 — Пример использования демонстрационного Web-интерфейса TerMine

### 1.3.5 Maui

Maui — система тематической классификации текстовых документов, работающая на основе методов обработки естественного языка и машинного обучения [2]. Схема функционирования системы приведена состоит из двух этапов работы: этапа первоначального построения и обучения модели, и этапа применения обученной модели к решению задачи тематической классификации текста.

Результаты тематической классификации, полученные при помощи Maui, могут рассматриваться в качестве тегов (меток) исходного текста. Без использования обученной модели, Maui функционирует как система автоматического извлечения ключевых фраз (в данной работе оценка системы Maui проводилась без использования обученной модели).

Система Maui является свободным кроссплатформенным программным обеспечением и распространяется на условиях лицензии GNU General Public License Version 3. При помощи Web-приложения maui-indexer [10] можно ознакомиться с основными возможностями системы (рисунок 1.5). В этом случае, ключевые слова и фразы перечислены в блоке “Keywords”.

В настоящий момент, Maui не способна обрабатывать русскоязычные тексты.

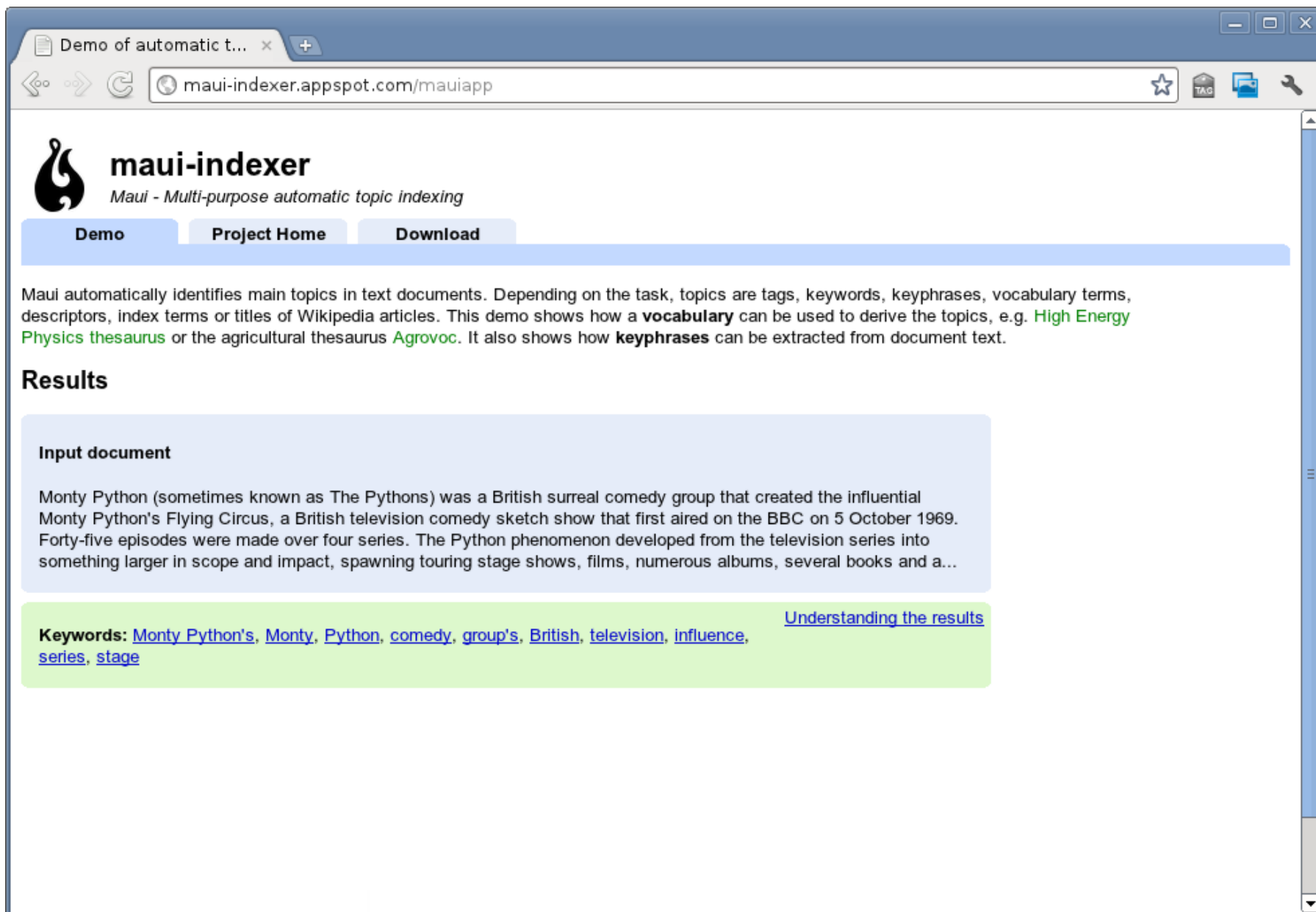


Рисунок 1.5 — Пример использования Web-приложения maui-indexer

### 1.3.6 TextAnalyst

TextAnalyst — инструмент для смыслового поиска информации и анализа содержания текстов [11], имеющий возможность выделения ключевых слов. Функционирование TextAnalyst основано на применении методов обработки естественного языка в сочетании с методами машинного обучения.

Пакет TextAnalyst доступен для ознакомительного использования в виде приложения для семейства операционных систем Microsoft® Windows® и способен обрабатывать тексты на русском языке.

Пример использования TextAnalyst приведён на рисунке 1.6. Ключевые слова исходного текста ключевые слова выделены зелёным цветом.

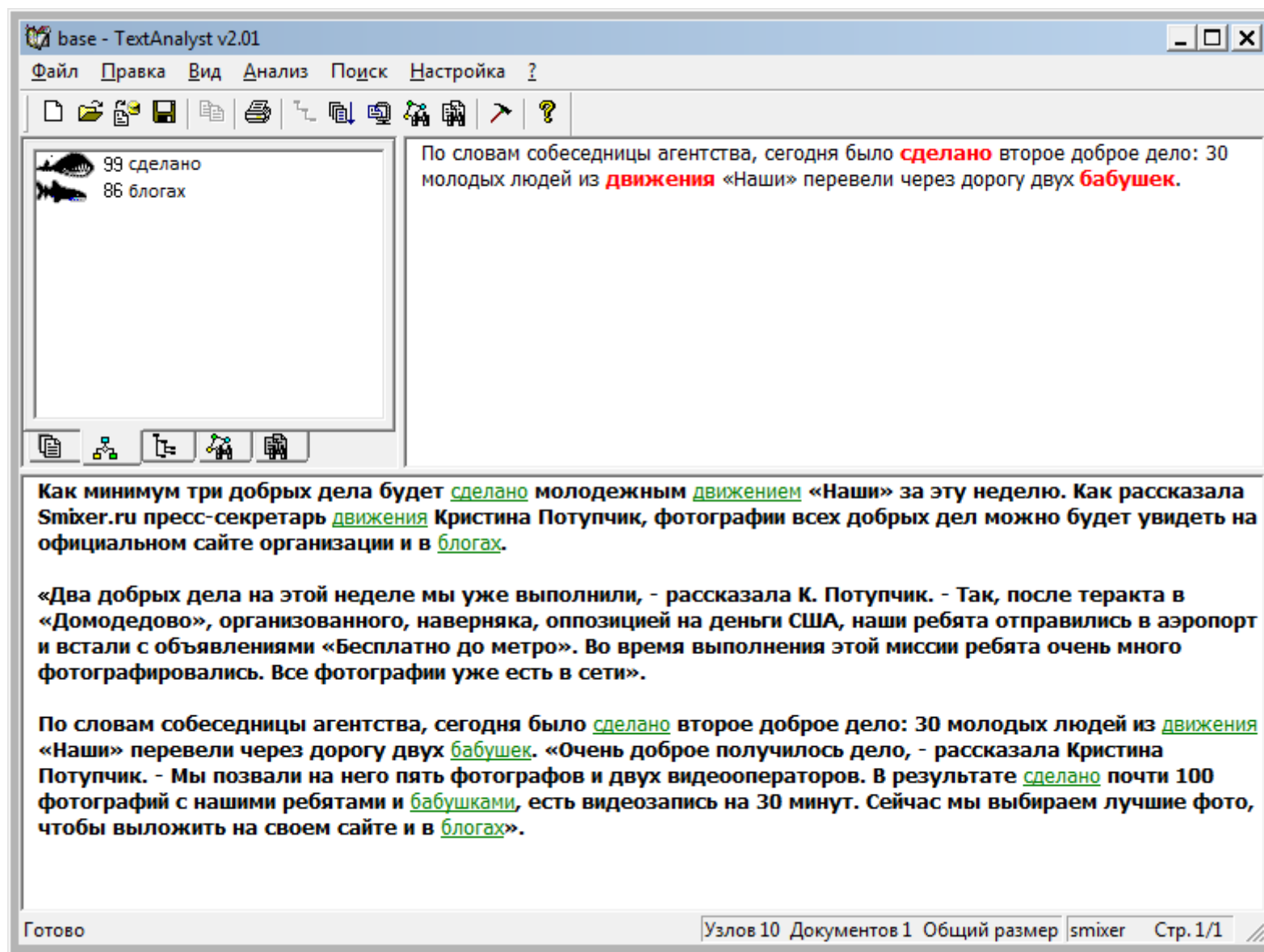


Рисунок 1.6 — Пример использования пакета TextAnalyst

### 1.3.7 АОТ

АОТ — проект, направленный на создание системы автоматического перевода «ДИАЛИНГ» [21]. В рамках проекта АОТ разработан комплекс инструментов автоматической обработки текста, в том числе графематический, морфологический и синтаксический анализаторы русского языка.

Все инструменты, разработанные в рамках проекта АОТ (в том числе и синтаксический анализатор) являются свободным кроссплатформенным программным обеспечением и распространяются на условиях лицензии GNU Lesser General Public License Version 2.1.

Пример использования анализатора приведён на рисунке 1.7. Именные группы, выделенные при помощи анализатора АОТ, можно принять в качестве терминов–кандидатов исходного текста [16].

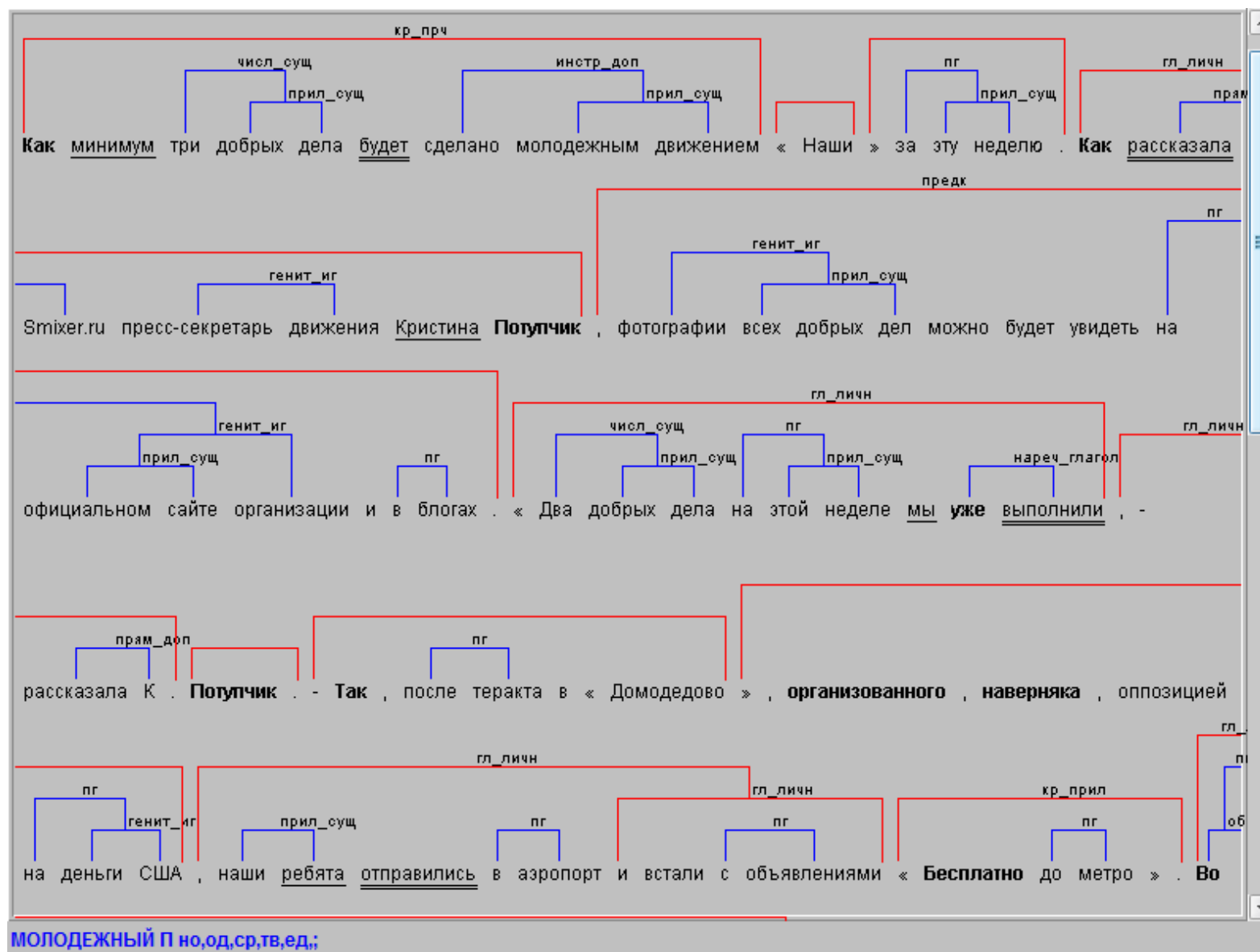


Рисунок 1.7 — Пример использования синтаксического анализатора АОТ

### 1.3.8 ContentAnalyzer

ContentAnalyzer — инструмент для анализа содержания тематических Web-страниц в реальном времени, выделения списков ключевых слов и словосочетаний, построения автореферата текста документа [13].

Функционирование ContentAnalyzer обеспечивается за счёт вычисления таких характеристик текста, как:

- частота термина/словосочетания в документе;
- отношение частоты к числу слов документа;
- вес термина в документе (с учётом частоты и весовых коэффициентов);
- вес термина к числу слов документа, и др.

Пакет ContentAnalyzer распространяется бесплатно, доступен для использования в виде приложения для семейства операционных систем Microsoft® Windows® и способен обрабатывать тексты как на русском, так и на английском языках.

Пример использования ContentAnalyzer приведён на рисунке 1.8: пакет рассчитал заявленные характеристики текста, выполнил извлечение ключевых слов и фраз, а также провёл автореферирование текста.

C:\Documents and Settings\eveel\My Documents\smixer.txt - ContentAnalyzer

Файл Настройки Справка

C:\Documents and Settings\eveel\My Documents\smixer.txt

Как минимум три добрых дела будет сделано молодежным движением «Наши» за эту неделю. Как рассказала Smixer.ru пресс-секретарь движения Кристина Потупчик, фотографии всех добрых дел можно будет увидеть на официальном сайте организации и в блогах.

«Два добрых дела на этой неделе мы уже выполнили, - рассказала К. Потупчик. - Так, после теракта в «Домодедово», организованного, наверняка, оппозицией на деньги США, наши ребята отправились в аэропорт и встали с объявлениями «Бесплатно до метро». Во время выполнения этой миссии ребята очень много фотографировались. Все фотографии уже есть в сети».

По словам собеседницы агентства, сегодня было сделано второе доброе дело: 30 молодых людей из движения «Наши» перевели через дорогу двух бабушек. «Очень доброе получилось дело, - рассказала Кристина Потупчик. - Мы позвали на него пять фотографов и двух видеооператоров. В результате сделано почти 100 фотографий с нашими ребятами и бабушками, есть видеозапись на 30 минут. Сейчас мы выбираем лучшие фото, чтобы выложить на своем сайте и в блогах».

Windows-1251 Download complete

Словооснова	С.	ЧА	ЧД	•ВА
добр		5	0,0333	2,50
дел		5	0,0333	2,50
фотограф		4	0,0266	2,00
потупчик		3	0,02	1,80
на		9	0,06	1,70
сделан		3	0,02	1,50
движен		3	0,02	1,50
рассказал		3	0,02	1,50

102 (67) / 150 (94) = 68,00% (71,28%)

Слово	С.	ЧА	ЧД	•ВА
Потупчик		3	0,02	1,80
добрых		3	0,02	1,50
сделано		3	0,02	1,50
рассказала		3	0,02	1,50
Наши		2	0,0133	1,20
Кристина		2	0,0133	1,20
дела		2	0,0133	1,00
движения		2	0,0133	1,00

94 (76) / 150 (110) = 62,67% (69,09%)

Словосочетание	Тип	ЧА	ВП	•ВСА	Поря...	Словоосновы
доброе дело		4	1,93	2,50	15	добр дел
движения Наши		2	0,49	1,60	34	движен на
Кристина Потупчик		2	1,03	1,50	57	крислин потупчик

8 (3) / 8 (3) = 100,00% (100,00%)

Реферрируемый текст

Предложение	Вес	ВСА	•Пор...	3...	ЧС	М
Как минимум три добрых дела будет сделано молодежным де...	12,20	0,94	0		13	
Как рассказала Smixer.ru пресс-секретарь движения Кристина...	19,30	0,92	1		21	
«Два добрых дела на этой неделе мы уже выполнили, рассказ...	10,30	0,94	2		11	
«Так, после теракта в «Домодедово», организованного, наvern...	11,85	0,52	3		23	
Во время выполнения этой миссии ребята очень много фотогр...	3,50	0,39	4		9	
Все фотографии уже есть в сети».	2,50	0,42	5		6	
По словам собеседницы агентства, сегодня было сделано вто...	15,30	0,73	6		21	
«Очень доброе получилось дело, рассказала Кристина Потупч...	10,00	1,43	7		7	
Мы позвали на него пять фотографов и двух видеооператоро...	6,30	0,70	8		9	

Слов: 149 / 150 = 99,33%

Авгореферат

Как минимум три добрых дела будет сделано молодежным движением «Наши» за эту неделю. Как рассказала Smixer.ru пресс-секретарь движения Кристина Потупчик, фотографии всех добрых дел можно будет увидеть на официальном сайте организации и в блогах. По словам собеседницы агентства, сегодня было сделано второе доброе дело: 30 молодых людей из движения «Наши» перевели через дорогу двух бабушек.

Предложений: 3, Слов: 55 / 150 = 36,67%

Рисунок 1.8 — Пример использования ContentAnalyzer

### 1.3.9 Семантическое зеркало

Семантическое зеркало — система тематической классификации Web-страниц [14], разработанная компанией «Ашманов и партнёры».

Сервис «Семантическое зеркало» обрабатывает текст Web-страницы и определяет её тему: анализирует слова, семантические связи между ними, выделяет самые важные термины. Темы определяются по рубрикатору, где к каждой рубрике приписано некоторое множество терминов.

Результат тематической классификации можно использовать в качестве списка ключевых фраз исходного текста или в качестве набора тегов (меток). Эту информацию можно использовать для показа контекстной рекламы и новостей на актуальную тему.

На сайте компании «Ашманов и партнёры» доступна демонстрационная версия «Семантического зеркала», имеющая лимит в 128 обращений с одного IP-адреса в сутки. Сервис обрабатывает тексты как на русском, так и на английском языках.

Пример использования сервиса приведён на рисунке 1.9: для каждого термина-кандидата вычислен его вес в исходном тексте. На основании этих данных корректно определена тема документа.

Демонстрация тех... x

www.ashmanov.com/tech/semantic/demo#

## 'Семантическое Зеркало'

Семантическое зеркало документа  
<http://koost.eveel.ru/nashi.html>

### Категории

Категория	Название	Вес
Society	Общество	54.2%

### Ключевые термины

Термин	Вес
Потупчик	12.4
Добрых делах	9.9
Фотографии	6.3
Делах	6.3
Добрых	6.1
Ребята	5
Кристина	4.3
Кристина Потупчик	2.1
Добрых дела сделано	2
Официальном сайте	1.1
Молодежным движением	1.1
Минимум три	1.1
Лучшие фото	1
Объявлениями Бесплатно	0.9
Время выполнения	0.9

Заполните все поля обязательные для заполнения.

Вас интересует: \*

Технологии / Поисковые сервисы

Ваше имя: \*

Телефон: \*\*

Электронная почта: \*\*

Сайт:

Комментарии, список ключевых слов, описание вашего бизнеса:

**ОТПРАВИТЬ**

\* — поля обязательные для заполнения

Рисунок 1.9 — Пример использования сервиса «Семантическое зеркало»

## 1.4 Формирование набора критериев

Необходимая система автоматического извлечения ключевых фраз из текста на естественном языке оценивается по следующим практически важным критериям:

— поддержка русского языка (“Р”):

0.0 — поддержка отсутствует;

1.0 — поддержка присутствует.

— качество результата по итогам экспертной оценки (“К”):

0.0 — минимальная оценка;

1.0 — максимальная оценка.

— доступность аналога (“Д”):

0.0 — использование аналога требует приобретения платной лицензии или временной подписки;

0.5 — существуют полноценные бесплатные версии аналога,  $\beta$ -версии или специальные версии для академических исследований;

1.0 — аналог распространяется как свободное программное обеспечение.

— независимость аналога от наличия онтологии заданной области знаний или специализированного тезауруса в процессе извлечения ключевых фраз (“О”):

0.0 — аналог спроектирован с целью использования специализированного тезауруса или онтологии области знаний в процессе выделения терминов;

1.0 — результат выделения терминов не зависит от наличия специализированного тезауруса или онтологии области знаний.

## 1.5 Сравнение аналогов

В результате обзора современных систем извлечения ключевых фраз из текста на естественном было найдено 9 аналогов. Для того, чтобы перейти к выбору прототипа, необходимо оценить каждый из аналогов в соответствии с критериями, обозначенными в 1.4.

Составим таблицу 1.1, в которой оценим аналоги по качественному уровню проявления критериев, выбранных в 1.4.

Таблица 1.1 — Сравнение существующих аналогов по заданным критериям.

№	Название аналога	Оценки по критериям				$\Sigma$
		Р	К	Д	О	
1	OpenCalais	0.0	0.8	0.5	0.0	1.3
2	Extractor	0.0	0.7	0.0	1.0	1.7
3	Yahoo! Term Extraction Web Service	0.0	0.6	0.5	1.0	2.1
4	TerMine	0.0	0.7	0.5	1.0	2.2
5	Maui	0.0	0.6	1.0	1.0	2.6
6	TextAnalyst	1.0	0.3	0.5	1.0	2.8
<b>7</b>	<b>АОТ</b>	1.0	0.4	1.0	1.0	<b>3.4</b>
8	ContentAnalyzer	1.0	0.6	0.5	1.0	3.1
9	Семантическое зеркало	1.0	0.5	0.5	1.0	3.0

## 1.6 Работа с прототипами

В результате попарного сравнения аналогов (таблица 1.1) по заданным критериям 1.4, целесообразным оказывается выбор аналога №7 (АОТ, выделено жирным) в качестве прототипа системы автоматического извлечения ключевых фраз из текста на естественном языке.

Структурная и алгоритмическая модели прототипа представлены на рисунках 1.10 и 1.11.

Компоненты, составляющие языковую модель (рисунок 1.10), — лингвистические процессоры, которые друг за другом обрабатывают входной текст. Вход одного процессора является выходом другого. Выделяются следующие компоненты:

- графематический анализатор;
- морфологический анализатор;
- синтаксический анализатор.

Графематический анализ — это начальный этап анализа естественного языкового текста, представленного в виде цепочки символов [22]. На этом этапе

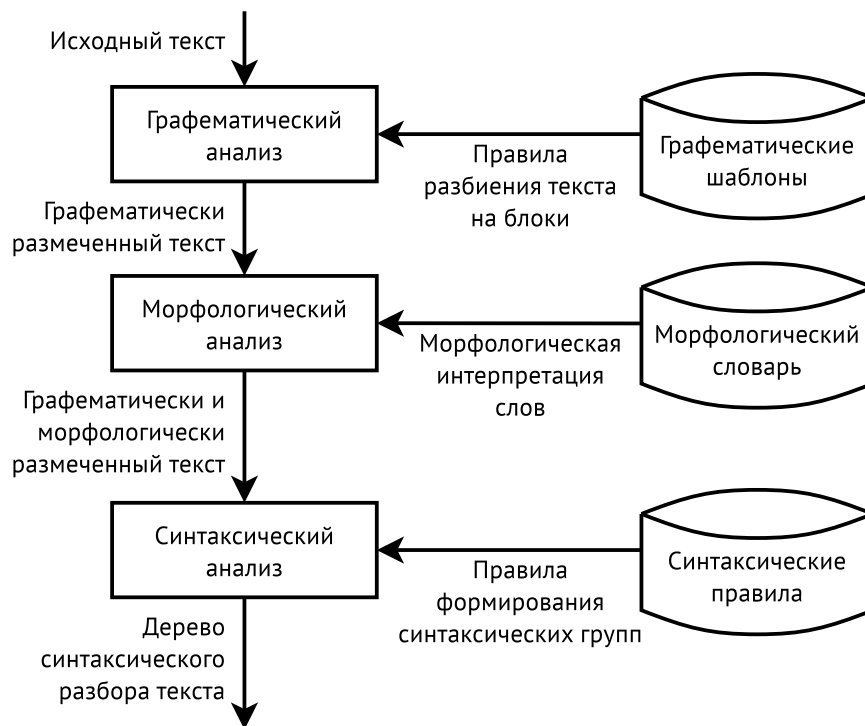


Рисунок 1.10 — Структурная модель прототипа системы автоматического извлечения ключевых фраз из текста на естественном языке

вырабатывается информация, необходимая для дальнейшей обработки морфологическим и синтаксическим анализаторами. В задачу графематического анализа входят: разделение входного текста на слова и разделители; выделение предложений из входного текста; выделение абзацев, заголовков, примечаний, и др. Подробное описание принципа функционирования графематического анализатора АОТ приведено в [22].

Задача морфологического анализатора — построение морфологической интерпретации слов входного текста [23]. Подробное описание морфологического анализатора АОТ приведено в [24]. Для каждого слова входного текста выдается множество морфологических интерпретаций следующего вида:

- морфологическая часть речи (например, род, число, падеж, и т. д.)
- лемма — каноническая форма лексемы (например, существительное в именительном падеже единственного числа, или глагол–инфинитив);
- множество наборов граммем — элементарных описателей, относящих словоформу к какому–либо морфологическому классу.

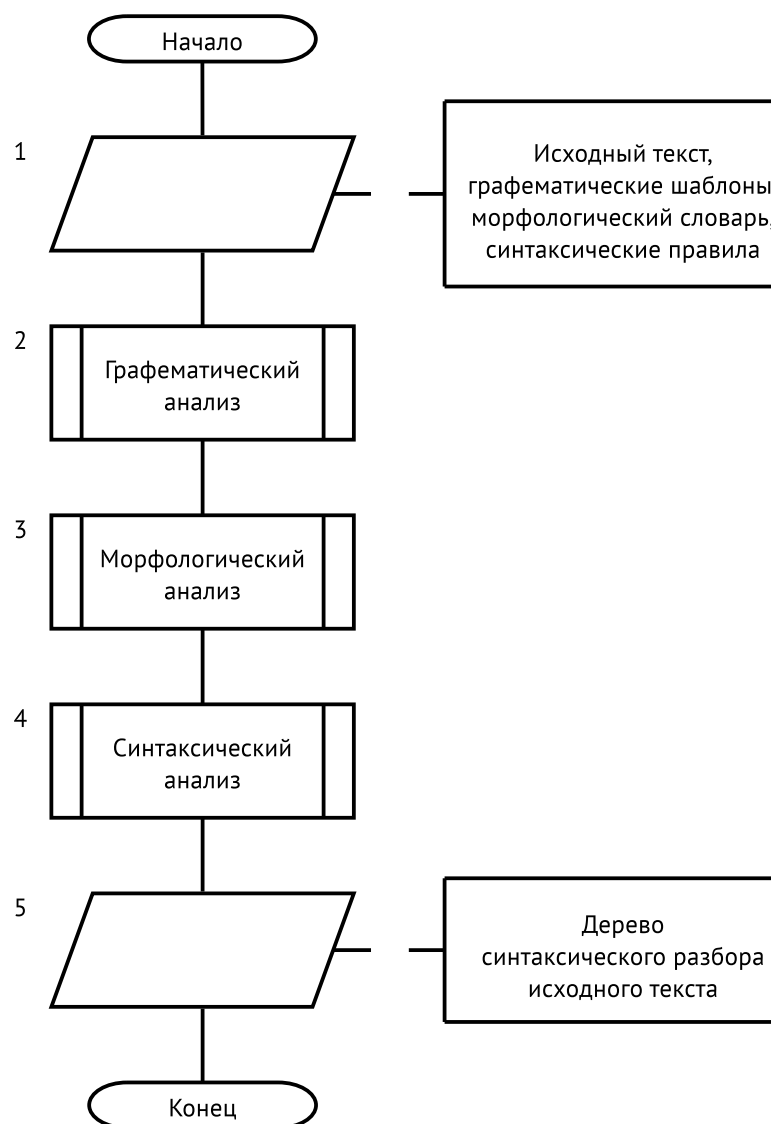


Рисунок 1.11 — Алгоритмическая модель прототипа системы автоматического извлечения ключевых фраз из текста на естественном языке

Цель синтаксического анализа — построение дерева зависимостей каждого предложения входного текста. Дерево представлено в виде ориентированного графа, вершинами которого являются синтаксические группы [25], построенные из слов входного текста, а дугами являются отношения между синтаксическими группами [23]. Группы строятся при помощи заранее определённых синтаксических правил [26]. Описание алгоритма синтаксического анализа АОТ приведено в [23].

Известно, что большинство терминов — это именные группы, что позволяет в качестве терминов-кандидатов рассматривать именные группы, выделенные с помощью синтаксического анализатора [16]. На рисунке 1.7 видно, что на основе слов входного текста получена древовидная структура, содержащая

различные синтаксические группы. Среди всех возможных групп, особый интерес для нашей задачи представляют именные группы (обозначены “прил\_сущ” и “гениг\_иг”):

- добрых дела;
- наши ребята;
- пресс-секретарь движения;
- и т. д.

Описание схемы функционирования прототипа наглядно представлено на рисунке 1.12 с учётом прохождения всех описанных уровней обработки текста, начиная с графематического анализа, заканчивая разбором предложений с формированием синтаксических групп.

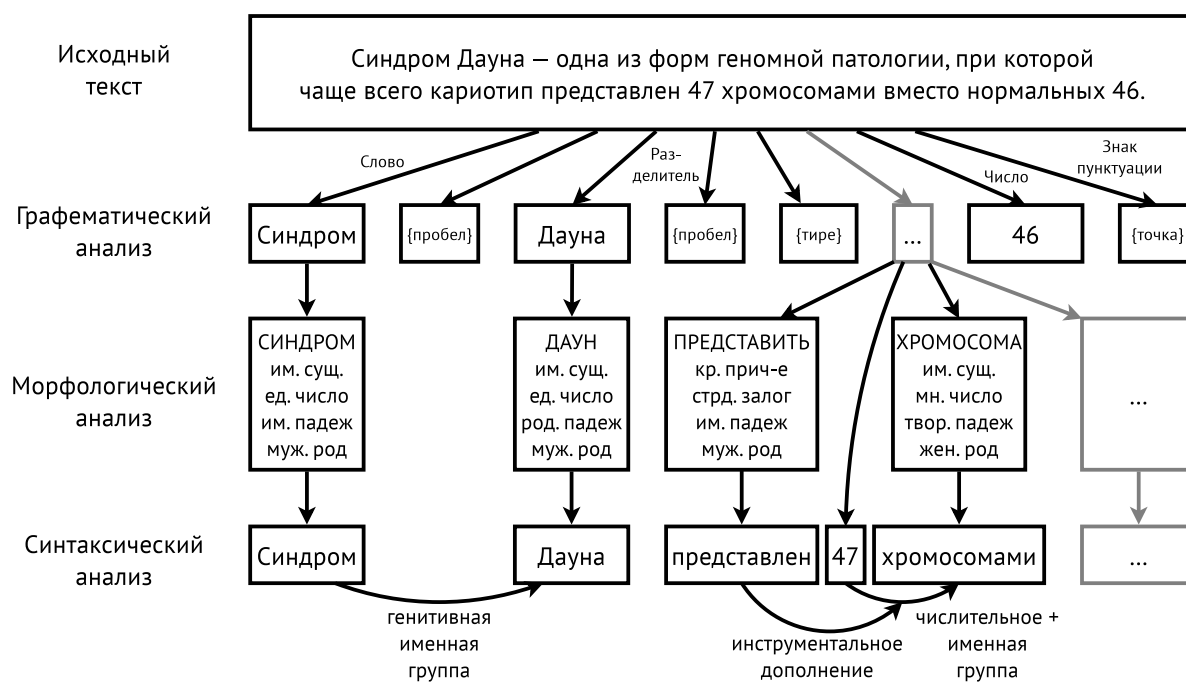


Рисунок 1.12 – Пример работы прототипа

### 1.6.1 Критика прототипа

Из структуры выбранного прототипа (рисунок 1.10), алгоритма его функционирования (рисунок 1.11) и результатов его сравнения с существующими аналогами (таблица 1.1) по критериям 1.4 очевидно, что применённый в прототипе метод распознавания многословных терминов обладает недостаточной точностью [16]. Необходима его доработка с целью явного определения

ключевых слов и фраз, наиболее адекватных исходному тексту, а также их ранжирования на основе статистического значения терминологичности.

Кроме того, стоит отметить, что качество работы современных морфологических анализаторов превосходит [27] морфологический анализатор АОТ [24].

### 1.6.2 Предлагаемое решение

В целях повышения адекватности результата извлечения ключевых фраз прототипом, предлагается внести в его структуру блок, вычисляющий статистическое значение терминологичности каждой именной группы, выделенной синтаксическим анализатором.

Из результатов попарного сравнения аналогов (таблица 1.1) видно, что аналог TerMine (1.3.4) получил наивысшую оценку качества результата работы среди систем, не использующих онтологию заданной области знаний в процессе выделения терминов.

Подобно системе TerMine, стоит воспользоваться статистическим методом C-value [4], что позволит сопоставить каждой извлечённой из текста именной группе значение терминологичности, вычисляемое по формуле:

$$C - value(a) = \begin{cases} \log_2 |a| \cdot f(a), & \text{если } a \text{ не вложен} \\ \log_2 |a| \cdot f(a) - \frac{1}{P(T_a)} \cdot \sum_{b \in T_a} f(b), & \text{если } a \text{ вложен,} \end{cases} \quad (1.1)$$

где  $a$  — кандидат в термины;

$|a|$  — длина словосочетания, измеряемая в количестве слов;

$f(a)$  — частотность  $a$ ;

$T_a$  — множество словосочетаний, которые содержат  $a$ ;

$P(T_a)$  — количество словосочетаний, содержащих  $a$ .

Легко видеть, что чем больше частота встречаемости термина–кандидата в тексте и чем выше его длина, тем больше его вес в исходном тексте. Однако если этот кандидат входит в большое количество других словосочетаний, то его вес уменьшается [4]. Путём сортировки списка кандидатов в термины по убыванию значения C-value можно получить список ключевых фраз, наиболее адекватных исходному тексту [16].

Таким образом, в структуру прототипа, изображенную на рисунке 1.10 целесообразно внести готовое решение 1-го ранга — блок явного выделения ключевых фраз на основе метода C-value.

Структурная и алгоритмическая модели блока выделения ключевых фраз (решения 1-го ранга) представлены на рисунках 1.13 и 1.14.

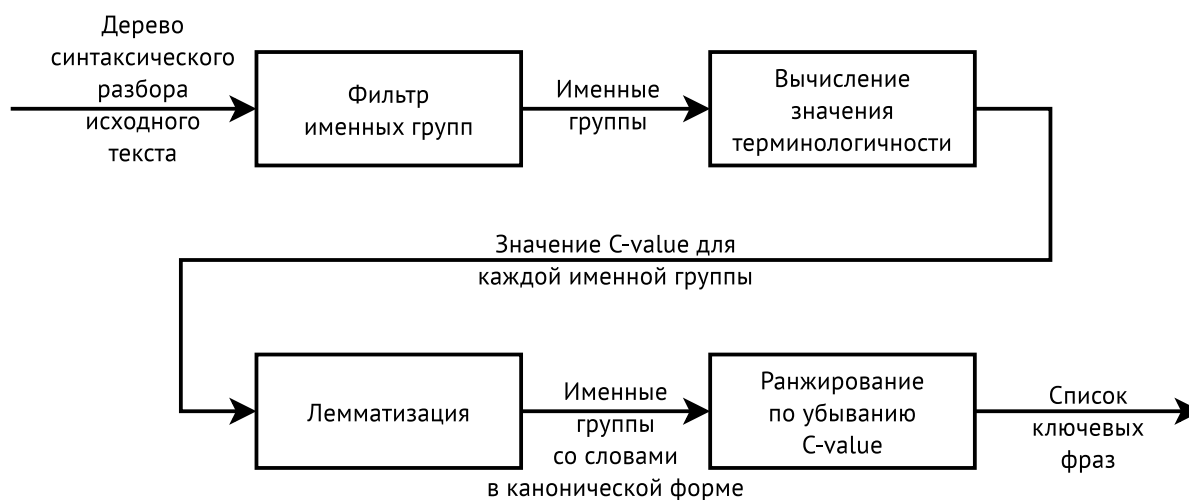


Рисунок 1.13 — Структурная модель блока выделения ключевых фраз — решения 1-го ранга

Также в качестве меры по повышению точности извлечения ключевых фраз из текста, необходимо модифицировать морфологический анализатор системы АОТ, что повысит качество предварительной обработки текста и положительно скажется на результате работы всей системы в целом.

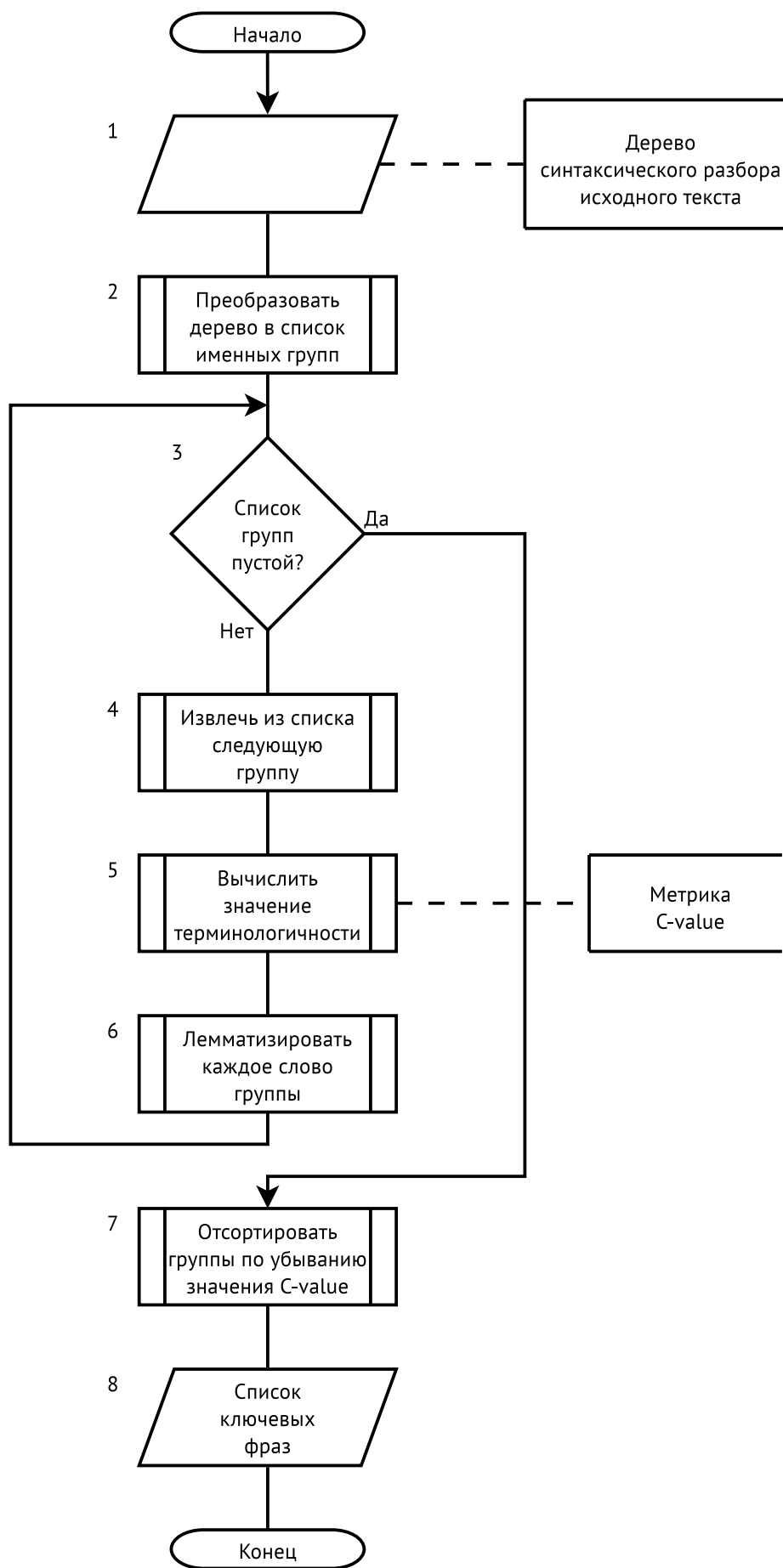


Рисунок 1.14 — Алгоритмическая модель блока выделения ключевых фраз — решения 1-го ранга

### 1.6.2.1 Обзор морфологических анализаторов

В качестве аналогов 1-го ранга по морфологическому анализу рассмотрим морфологический анализатор АОТ и сравним его с другими известными [27] морфологическими анализаторами:

- *mystem* [28, 29], разработчик — компания «Яндекс»;
- *Snowball* [30, 31], разработчик — Dr. M. Porter;
- *myaso* [32], разработчик — Д. Усталов;
- *rumorphy* [33], разработчик — М. Коробов;
- *TreeTagger* [20, 34], разработчик — H. Schmid;
- *TnT* [35, 36], разработчик — T. Brants.

В качестве критериев их сравнения выберем:

- поддержка русского языка (“Р”):
  - 0.0 — поддержка отсутствует;
  - 1.0 — поддержка присутствует.
- возможность определения части речи (англ. *POS-tagging*) и грамматических характеристик слова (“Ч”):
  - 0.0 — часть речи и морфологические шаблоны не определяются;
  - 0.5 — выполняется определение только части речи слова;
  - 1.0 — определяется часть речи и грамматические характеристики слова.
- возможность выделения основы слова (англ. *stemming*) (“О”):
  - 0.0 — основа слова не выделяется;
  - 1.0 — основа слова выделяется.
- качество анализа по итогам экспертной оценки (“К”):
  - 0.0 — минимальная оценка;
  - 1.0 — максимальная оценка.
- доступность аналога (“Д”):
  - 0.0 — использование аналога требует приобретения платной лицензии или временной подписки;

0.5 — существуют полноценные бесплатные версии аналога,  $\beta$ -версии или специальные версии для академических исследований;

1.0 — аналог распространяется как свободное программное обеспечение.

Таблица 1.2 — Сравнение существующих аналогов 1-го ранга по заданным критериям.

№	Название аналога	Оценки по критериям					$\Sigma$
		Р	Ч	О	К	Д	
1	АОТ	1.0	1.0	1.0	0.6	1.0	4.6
2	mystem	1.0	1.0	1.0	0.8	0.5	4.3
3	Snowball	1.0	0.0	1.0	0.6	1.0	3.6
<b>4</b>	<b>myaso</b>	1.0	1.0	1.0	0.7	1.0	<b>4.7</b>
5	rumorphy	1.0	1.0	1.0	0.5	1.0	4.5
6	TreeTagger	1.0	1.0	1.0	0.7	0.5	4.2
7	TnT	1.0	1.0	0.0	0.7	0.5	3.2

В результате сравнения аналогов (таблица 1.2), целесообразно использовать аналог №4 (myaso, выделено жирным) в качестве морфологического анализатора — готового решения 1-го ранга.

Структурная и алгоритмическая модели морфологического анализатора (решения 1-го ранга) представлены на рисунках 1.15 и 1.16.

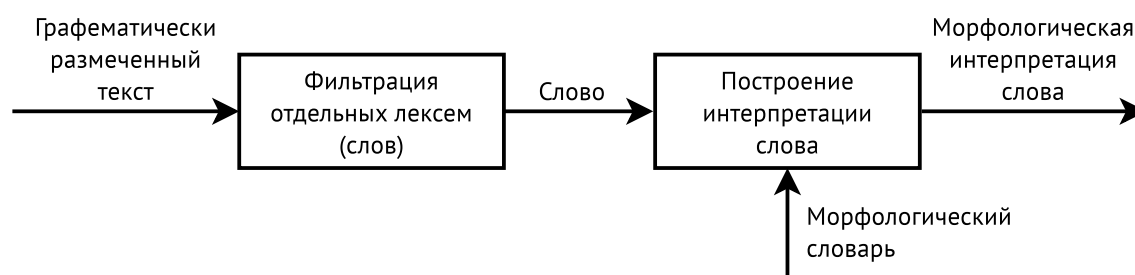


Рисунок 1.15 — Структурная модель морфологического анализатора — решения 1-го ранга

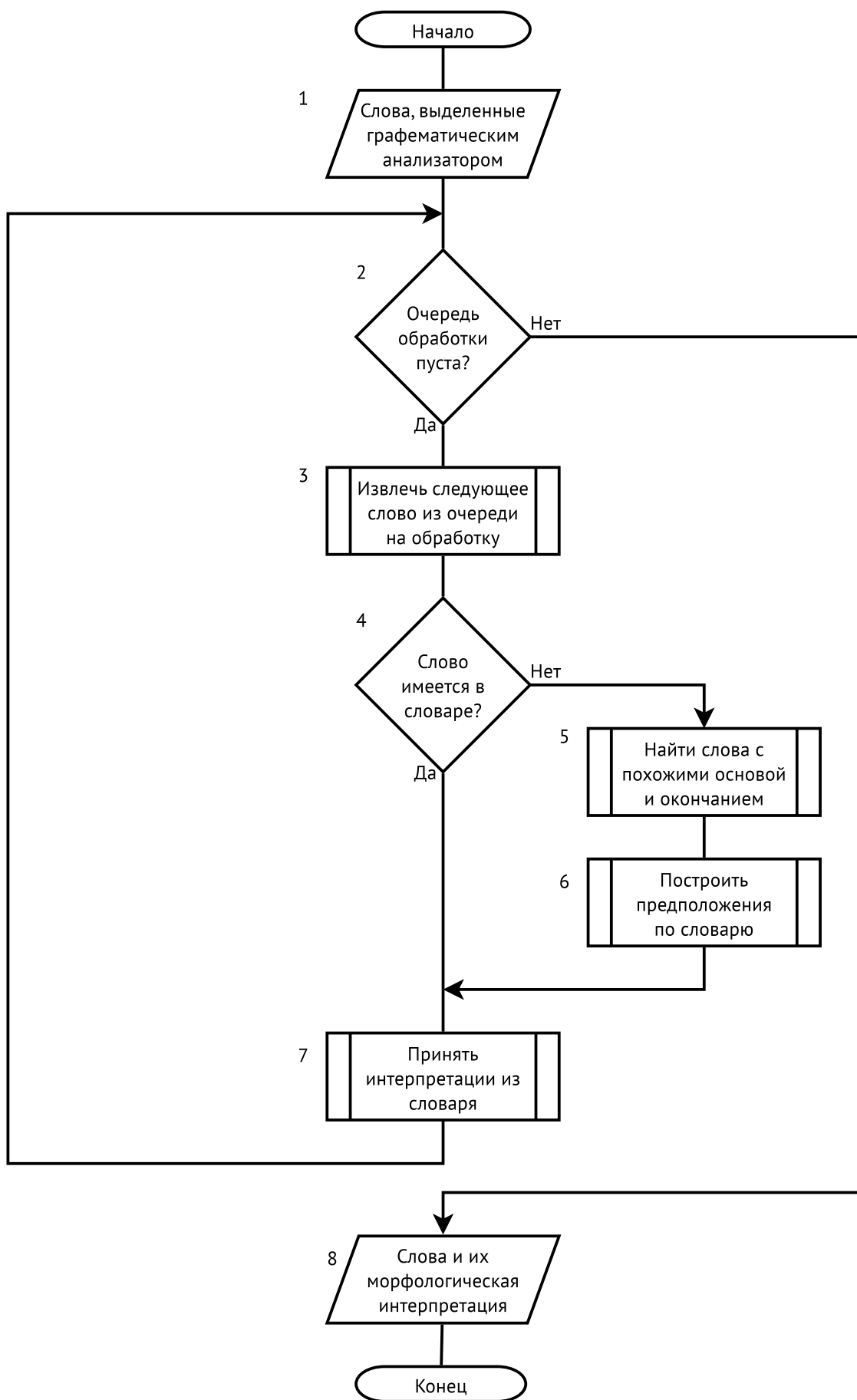


Рисунок 1.16 — Алгоритмическая модель морфологического анализатора — решения 1-го ранга

## 1.7 Результаты и выводы

В ходе выполнения литературно–аналитического обзора по теме выпускной квалификационной работы:

- приведена технология 1.2 поиска информации в экспертных, электронных и библиографических источниках;
- на основе технологии поиска информации 1.2 выбрано 48 источников литературы по теме выпускной квалификационной работы бакалавра;
- найдено 9 аналогов системы автоматического извлечения ключевых фраз из текста на естественном языке;
- определён пакет эмпирических экспертных критериев 1.4 для оценивания найденных аналогов;
- составлена таблица 1.1 попарного сравнения аналогов по критериям, обозначенным в 1.4;
- на основе результатов сравнения аналогов (таблица 1.1) выбран прототип системы автоматического извлечения ключевых фраз из текста на естественном языке;
- проведён критический анализ прототипа и предложены пути его развития.

## 2 Модель предлагаемого решения

### 2.1 Содержательная модель

Система автоматического извлечения ключевых фраз из текста на естественном языке выполняет выделение ключевых фраз из текста на естественном языке с применением графематических шаблонов, морфологического словаря (лексикона) и синтаксических правил. Эти данные определены предварительно и хранятся в базе данных.

Текст обрабатывается графематическим анализатором, который вырабатывает информацию о разделении текста на абзацы, предложения и отдельные слова, необходимую для дальнейшей обработки.

Каждое слово, выделенное графематическим анализатором, подвергается морфологическому анализу с целью построения морфологической интерпретации (часть речи, форма, и т. д.), определения основы слова и формирования леммы (канонической формы лексемы).

На основе имеющейся графематической и морфологической интерпретации текста, выполняется построение и наполнение синтаксических групп, и выявление отношений между ними.

Ключевые фразы выделяются из именных групп, сформированных синтаксическим анализатором при помощи статистического метода C-value, поощряющего существование в тексте ключевых фраз, не входящих в состав других, более длинных.

### 2.2 Концептуальная модель

#### 2.2.1 Общая модель

Система автоматического извлечения ключевых фраз из текста на естественном языке — программный комплекс, выполняющий *функции* статистического выделения ключевых фраз из текста на естественном языке *путём* графематического, морфологического, синтаксического анализа текста *на основе* определённых графематических шаблонов, морфологических словарей, синтаксических правил и метрики C-value, *направленный* на автоматизацию процедуры извлечения ключевых фраз *с целью* формирования списка терминов по тексту.

### 2.2.2 Базово–уровневая модель

**Функции:** статистическое выделение ключевых фраз из текста (определение списка терминов–кандидатов из именных групп, подсчёт частоты появления в тексте каждого термина–кандидата, вычисление значения терминологичности для каждого термина–кандидата из списка).

**Пути реализации функций:** построение графематической интерпретации входного текста (разбиение текста на абзацы, предложения, подпредложения и отдельные слова, определение языка слов); построение морфологической интерпретации слов входного текста (выделение основ слов, определение частей речи и соответствующих наборов граммем); синтаксический разбор входного текста (построение синтаксических групп на каждом морфологическом варианте каждого фрагмента предложения, выявление связей между группами).

**Основы выполнения функций:** определённые графематические шаблоны, морфологические словари и синтаксические правила, метрика C-value.

**Направление реализации функций:** автоматизация процедуры извлечения ключевых фраз из текста.

**Цель:** формирование списка терминов по тексту.

### 2.2.3 Модификационная модель

**Функции:** статистическое выделение ключевых фраз из текста (определение списка терминов–кандидатов из именных групп, подсчёт частоты появления в тексте каждого термина–кандидата, нахождение терминов–кандидатов в составе других терминов–кандидатов, подсчёт длины каждого термина–кандидата, вычисление значения терминологичности для каждого термина–кандидата из списка).

**Пути реализации функций:** построение графематической интерпретации входного текста (разбиение текста на абзацы, предложения, подпредложения и отдельные слова при помощи детерминированного конечного автомата; определение языка каждого слова при помощи регулярных выражений); построение морфологической интерпретации слов входного текста (поиск оптимального разбиения каждого слова на суффикс и основу слова, определение правил внутренней флексии для словоформ, определение частей речи и соответ-

ствующих наборов граммем по морфологическому словарю); синтаксический разбор входного текста (построение синтаксических групп на каждом морфологическом варианте каждого фрагмента предложения, выявление связей между группами, идентификация вида каждой связи).

**Основы выполнения функций:** определённые графематические шаблоны, морфологические словари и синтаксические правила, метрика C-value.

**Направление реализации функций:** автоматизация процедуры извлечения ключевых фраз из текста.

**Цель:** формирование списка списка терминов по тексту.

### 2.3 Структурная модель

Структурная модель предлагаемого решения приведена на рисунке 2.1. В структуру прототипа внесён блок выделения ключевых фраз (выделено серым цветом), рассматриваемый как готовое решение 1-го ранга, изображённое на рисунке 1.13.

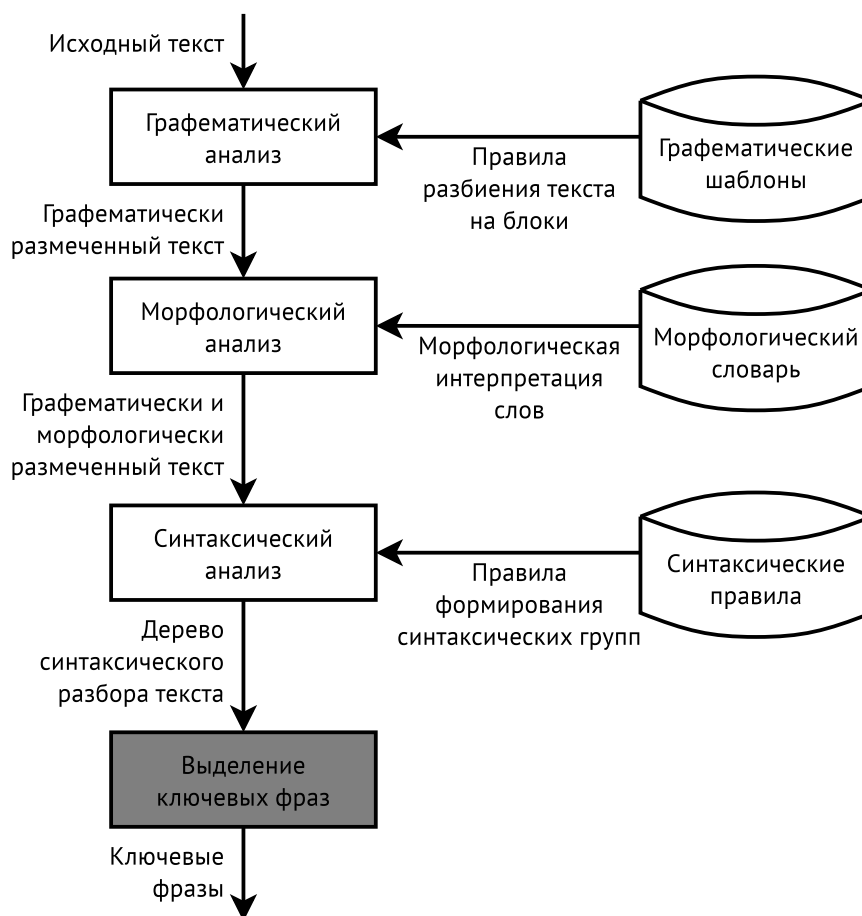


Рисунок 2.1 — Структурная модель предлагаемого решения

## 2.4 Функционально–структурная модель

Функционально–структурная модель предлагаемой системы автоматического извлечения ключевых фраз из текста на естественном языке построена при помощи пакета Computer Associates ERwin Process Modeler 7.3 с применением методологии проектирования SADT IDEF0 и приведена в приложении А.

На рисунке А.1 показано функционирование системы с точки зрения разработчика (0-й уровень, контекстная диаграмма). При последующей декомпозиции получается функционально–структурная модель 1-го уровня (рисунок А.2).

Блок №4: «Выделить ключевые фразы» отсутствует в прототипе и был внесён с целью преодоления его недостатков, обозначенных в 1.6.1. Декомпозиция узла А4, соответствующего данному блоку, представлена на рисунке А.3.

## 2.5 Алгоритмическая модель

Алгоритмическая модель предлагаемого решения приведена на рисунке 2.2. В алгоритм функционирования прототипа внесено готовое решение 1-го ранга — процедура выделения ключевых фраз (обозначено серым цветом), алгоритм функционирования которой изображён на рисунке 1.14.

## 2.6 Результаты и выводы

Были построены следующие модели предлагаемого решения:

- содержательная модель;
- концептуальные модели: общая, базово–уровневая и модификационная;
- структурная модель;
- функционально–структурная модель;
- алгоритмическая модель.

Разработанный пакет моделей системы автоматического извлечения ключевых фраз из текста на естественном языке даёт представление о принципах её функционирования, концептуальном устройстве, а также структурном составе.

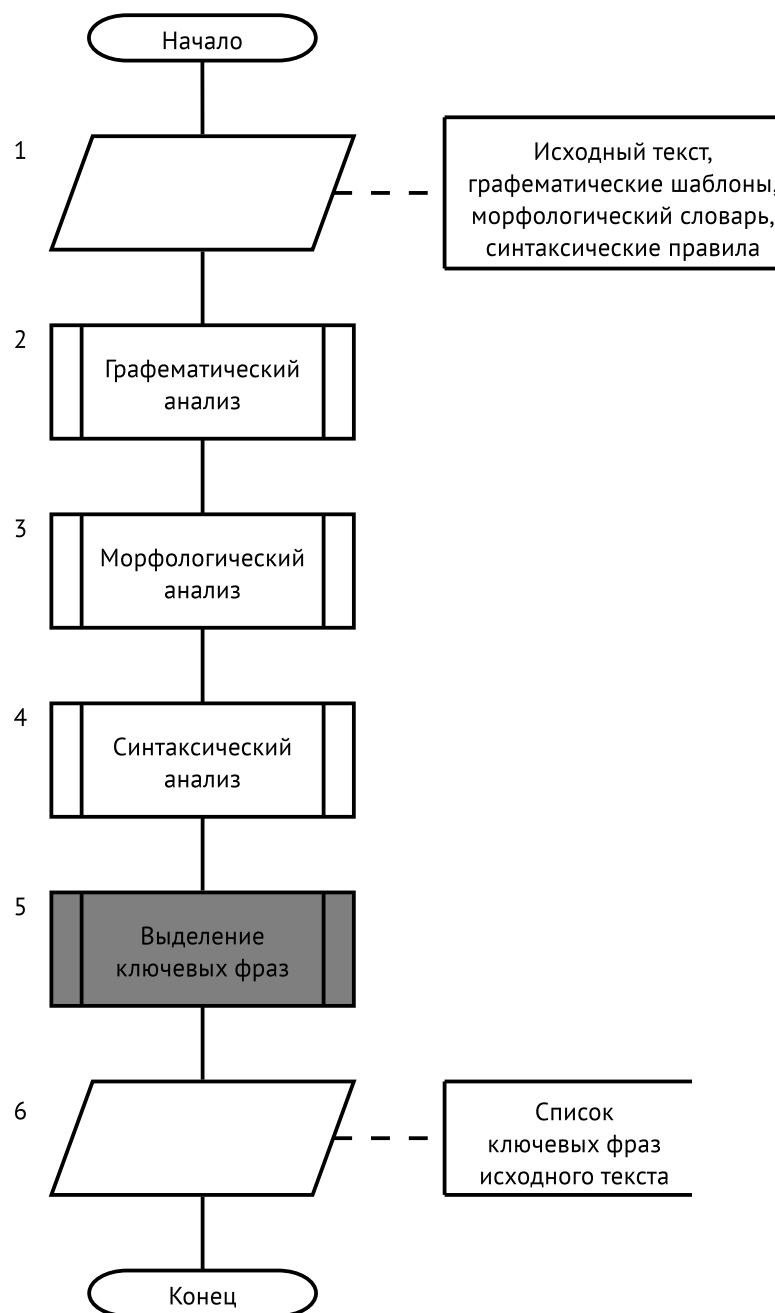


Рисунок 2.2 — Алгоритмическая модель предлагаемого решения

Созданный пакет моделей позволяет провести проектирование системы автоматического извлечения ключевых фраз из текста на естественном языке, устранив основные недостатки прототипа, обозначенные в 1.6.1.

## **3 Проектирование предлагаемого решения**

### **3.1 Внешнее проектирование**

В рамках проведения внешнего проектирования разработано Техническое задание, которое приведено в приложении Б.

### **3.2 Внутреннее проектирование**

Внутреннее проектирование проводилось на основании моделей, разработанных и описанных в главе 2. В ходе внутреннего проектирования решались задачи, сформулированные в Техническом задании.

Для создания системы автоматического извлечения ключевых фраз из текста на естественном языке была выбрана концепция мультипарадигменного программирования, в частности объектно-ориентированного и функционального программирования.

В качестве основной платформы для разработки выбран Rubinius — перспективная реализация языка программирования Ruby [37] на основе виртуальной машины LLVM [38]. В соответствии с современными требованиями и тенденциями, предлагаемое решение построено на основе архитектуры REST [17], а также стандартных протоколов: JSON [39] и XML [18].

### **3.3 Результаты и выводы**

В ходе выполнения внешнего и внутреннего проектирования системы автоматического извлечения ключевых фраз из текста на естественном языке получены следующие результаты:

- разработано Техническое задание на создание системы;
- определён состав инструментальных средств для создания системы.

На основе результатов проектирования можно сделать следующий вывод: разработанный проект системы соответствует Техническому заданию.

## 4 Инженерная реализация и эксплуатация предлагаемого решения

### 4.1 Требования к средствам обеспечения

Основываясь на требованиях к аппаратному обеспечению, указанных в Техническом задании (см. приложение Б), для инженерной реализации, тестирования и эксплуатации системы автоматического извлечения ключевых фраз из текста на естественном языке, характеристики конфигурации *сервера* не должны быть ниже следующих:

- центральный процессор Intel® Core® Duo или аналогичный по производительности;
- оперативная память объёмом 2048 мегабайт;
- жёсткий диск объёмом 160 гигабайт;
- сетевая карта, работающая в полном дуплексе со скоростью не менее 100 мегабит в секунду;
- операционная система: любая операционная система, на которой возможен запуск виртуальной машины Rubinius (например, Red Hat® Enterprise Linux® Server 6 [40]).

Характеристики конфигурации *рабочей станции пользователя* для работы с системой не должны быть ниже следующих:

- центральный процессор Intel® Pentium® 4 или аналогичный по производительности;
- оперативная память объёмом 1024 мегабайт;
- жёсткий диск объёмом 160 гигабайт;
- сетевая карта, работающая в полном дуплексе со скоростью не менее 10 мегабит в секунду;
- периферийные устройства:
  - монитор;
  - клавиатура;
  - мышь.

— операционная система: на выбор пользователя — Microsoft® Windows®, GNU/Linux, Mac OS X, Oracle® Solaris™, и др.

Инженерная реализация системы автоматического извлечения ключевых фраз из текста на естественном языке выполнена на мультипарадигменном языке программирования Ruby с использованием платформы Rubinius.

## 4.2 Экранные формы

Экранные формы разработанной системы автоматического извлечения ключевых фраз из текста на естественном языке созданы в соответствии со стандартом HTML5 [41].

При обращении к системе, пользователю предлагается предоставить исходный текст для обработки и выбрать формат, в котором будет представлен результат работы системы: HTML, JSON или XML (рисунок 4.1).



Рисунок 4.1 — Главная страница Web-интерфейса Tesuqk — системы автоматического извлечения ключевых фраз из текста на естественном языке

После выполнения запроса на обработку текста при выбранном формате представления результатов HTML (рисунок 4.2), система оформляет результат в виде таблицы, состоящей из трёх колонок:

- № — номер ключевой фразы, в соответствии с её значением терминологичности;
- Ключевая фраза — именная группа, выделенная в качестве ключевой фразы исходного текста;
- C-value — значение терминологичности, вычисленное в соответствии с описанным в 1.6.2.

### **4.3 Результаты и выводы**

В ходе инженерной реализации и эксплуатации разработанной системы автоматического извлечения ключевых фраз из текста на естественном языке, получены следующие результаты:

- программно реализована система автоматического извлечения ключевых фраз из текста на естественном языке;
- проведены эксперименты на работоспособность системы.

# Tesuçk

Tesuçk

Что это?!

API

Система **Tesuçk** находится в состоянии  $\alpha$ -тестирования.

Выделенные ключевые слова и фразы

№	Ключевая фраза	C-value
1	официальном сайте организации	4.858
2	молодежным движением	4.322
3	словам собеседницы	4.170
4	второе доброе дело	4.170
5	секретарь движения	4.170
6	время выполнения	4.000
7	видеооператоров	3.907
8	после теракта	3.700
9	молодых людей	3.700
10	объявлениями	3.585
11	видеозапись	3.459
12	результате	3.322
13	оппозицией	3.322
14	фотографии	3.322
15	фотографов	3.322
16	домодедово	3.322
17	агентства	3.170
18	аэропорт	3.000
19	движения	3.000
20	крестина	3.000
21	минимум	2.807
22	бабушек	2.807
23	деньги	2.585
24	ребята	2.585
25	дорогу	2.585
26	миссии	2.585
27	неделю	2.585
28	пресс	2.322
29	минут	2.322
30	сайте	2.322
31	метро	2.322
32	фото	2.000
33	сети	2.000
34	дела	2.000
35	сша	1.585
36	уже	1.585
37	из	1.000

Выделено кандидатов в термины: 37.

Tesuçk v0.1alpha © [Дмитрий Усталов](#) ♥ при поддержке [ИММ УрО РАН](#).

Рисунок 4.2 — Web-интерфейс представления результатов работы Tesuçk — системы автоматического извлечения ключевых фраз из текста на естественном языке

## Заключение

В процессе построения системы автоматического извлечения ключевых фраз из текста на естественном языке был проведён обзор аналогов, выбран прототип, проведён его критический анализ, указаны недостатки и пути их устранения. Это позволило сформулировать цели и задачи для создания пакета моделей.

На этапе построения моделей получены содержательная, концептуальная, структурная, функционально-структурная и алгоритмическая модели предлагаемого решения.

Данные модели позволили уточнить функцию и структуру системы автоматического извлечения ключевых фраз из текста на естественном языке, найти пути устранения недостатков прототипа, а также составить Техническое задание на программный продукт.

В ходе внутреннего проектирования решены задачи, сформулированные в Техническом задании, исследована структура и взаимосвязь составляющих системы автоматического извлечения ключевых фраз из текста на естественном языке.

Реализован Tesuck — система автоматического извлечения ключевых фраз из текста на естественном языке. Web-интерфейс системы доступ по адресу <http://tesuck.eveel.ru/>.

### Дальнейшая работа

Сегодня в ведущих иностранных научных сообществах наблюдается большой интерес к системам синтеза изображения по тексту (*англ.* TTP — Text-to-Picture) и их приложениям [42–44]. Использование таких систем целесообразно, когда применение традиционного текстового человеко-машинного интерфейса невозможно или недостаточно эффективно [43].

Известно, что иллюстрация делает текст нагляднее и проще в восприятии [44]. Этот факт повсеместно используется в различных дидактических системах, например при обучении родному языку детей дошкольного возраста (или обучении иностранному языку взрослых людей). Наличие иллюстраций в

тексте способствует пополнению словарного запаса обучающегося и развитию его ассоциативного мышления [45].

При помощи ТТР-систем возможно автоматически синтезировать изображения, отражающие ключевые фрагменты текста, что позволяет широко применять их в компьютерных обучающих системах [44].

Важнейшей задачей является медицинская реабилитация глухих людей, или имеющих черепно-мозговые травмы, или различные умственные нарушения [42] (например, синдром Дауна, и др.). При помощи систем синтеза изображения по тексту возможно радикально упростить представление учебной информации, тем самым обеспечив врачей и педагогов качественным дидактическим материалом [43, 45], сделав процесс обучения по-настоящему наглядным.

Функционирование ТТР-систем основано [44] на предварительном выделении *ключевых слов и фраз* из исходного текста, и последующей его обработкой при помощи методов компьютерной лингвистики [46], машинного обучения [47], компьютерного зрения [48], и др.

Несмотря на достоинства и возможности ТТР-систем, не обнаружено ни одной ТТР-системы, способной работать с русскоязычными текстами. Следовательно, целесообразно создание системы синтеза изображения по тексту на основе современной системы автоматического извлечения ключевых фраз.

На основе системы Tesuçk, созданной в рамках данной выпускной квалификационной работы, возможно построение современной ТТР-системы, способной обрабатывать русскоязычные тексты.

### **Благодарности**

Автор благодарит С. И. Кумкова за ценные замечания по содержанию работы.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Браславский П., Соколов Е.* Сравнение четырёх методов автоматического извлечения двухсловных терминов из текста // Труды международной конференции «Диалог 2006». — Бекасово: 2006. — С. 7.
2. *Medelyan O.* Human-competitive automatic topic indexing: Ph.D. thesis / The University of Waikato. — 2009. — 244 pp.
3. *Turney P. D.* Learning Algorithms for Keyphrase Extraction // *Information Retrieval — INRT 34-99*. — 2000. — P. 46.
4. *Frantzi K., Ananiadou S., Mima H.* Automatic recognition of multi-word terms: the C-value/NC-value method // *International Journal on Digital Libraries*. — 2000. — Vol. 3. — Pp. 115–130.
5. KEA: Practical automatic keyphrase extraction / I.H. Witten, G.W. Paynter, E. Frank et al. // *Proceedings of the fourth ACM conference on Digital libraries*. — 1999. — Pp. 254–255.
6. Home | OpenCalais [Электронный ресурс] // [сайт]. URL: <http://www.opencalais.com> (дата обращения: 29.03.2011).
7. Extractor Live Content Demonstration [Электронный ресурс] // [сайт]. URL: <http://www.extractorlive.com> (дата обращения: 02.02.2011).
8. Term Extraction Web Service - YDN [Электронный ресурс] // [сайт]. URL: <http://developer.yahoo.com/search/content/V1/termExtraction.html> (дата обращения: 11.01.2011).
9. TerMine Web Demonstrator [Электронный ресурс] // [сайт]. URL: <http://www.nactem.ac.uk/software/terminer> (дата обращения: 29.12.2010).
10. Maui - Multi-purpose automatic topic indexing [Электронный ресурс] // [сайт]. URL: <http://code.google.com/p/maui-indexer> (дата обращения: 30.03.2011).
11. New approach to text analysis [Электронный ресурс] // Microsystems, Ltd. [сайт]. URL: <http://analyst.ru/index.php?lang=eng&dir=content/tech/&id=approach> (дата обращения: 30.03.2011).
12. Проект АОТ: Синтаксический анализ [Электронный ресурс] // [сайт]. URL: <http://www.aot.ru/docs/synan.html> (дата обращения: 29.03.2011).

13. Content Analyzer v0.52 [Электронный ресурс] // [сайт]. URL: <http://www.rvsn2.narod.ru/soft51.htm> (дата обращения: 29.03.2011).
14. Технология “Семантическое зеркало” [Электронный ресурс] // Ашманов и партнеры [сайт]. URL: <http://www.ashmanov.com/tech/semantic> (дата обращения: 22.04.2011).
15. *Браславский П., Соколов Е.* Автоматическое извлечение терминологии с использованием поисковых машин Интернета // Труды международной конференции «Диалог 2007». — Бекасово: 2007. — С. 6.
16. *Браславский П., Соколов Е.* Сравнение пяти методов извлечения терминов произвольной длины // Труды международной конференции «Диалог 2008». — Бекасово: 2008. — С. 8.
17. *Fielding R. T.* Architectural Styles and the Design of Network-based Software Architectures: Ph.D. thesis. — The University of California, 2000. — 180 pp.
18. Extensible Markup Language (XML) 1.0 (Fifth Edition) [Электронный ресурс] // W3C Recommendation [сайт]. URL: <http://www.w3.org/TR/REC-xml/> (дата обращения: 27.04.2011).
19. *Brownlee J.* Clever Algorithms: Nature-Inspired Programming Recipes. — 2011.
20. TreeTagger - a language independent part-of-speech tagger [Электронный ресурс] // [сайт]. URL: <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html> (дата обращения: 22.04.2011).
21. Проект АОТ: Главная [Электронный ресурс] // [сайт]. URL: <http://aot.ru> (дата обращения: 29.03.2011).
22. Проект АОТ: Графематика [Электронный ресурс] // [сайт]. URL: <http://www.aot.ru/docs/graphan.html> (дата обращения: 29.03.2011).
23. *Ножов И. М.* Морфологическая и синтаксическая обработка текста: модели и программы: Дис. . . канд. техн. наук. — М., 2003. — 140 с.
24. *Сокирко А.* Морфологические модули на сайте [www.aot.ru](http://www.aot.ru) // Труды международной конференции «Диалог 2004». — Бекасово: 2004. — С. 7.
25. *Levine R., Meurers D.* Head-Driven Phrase Structure Grammar: Linguistic Approach, Formal Foundations, and Computational Realization // *Encyclo-*

*pedia of Language and Linguistics, Second Edition.* — 2006.

26. *Гладкий А. В.* Синтаксические структуры естественного языка в автоматизированных системах общения / Под ред. Д. А. Поспелов. Проблемы искусственного интеллекта. — М.: Наука, 1985. — С. 144.

27. Оценка методов автоматического анализа текста: морфологические парсеры русского языка / О. Ляшевская, И. Астафьева, А. Бонч-Осмоловская и др. // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог 2010». — Бекасово: 2010.

28. О программе *mystem* [Электронный ресурс] // Компания Яндекс [сайт]. URL: <http://company.yandex.ru/technology/mystem/> (дата обращения: 22.04.2011).

29. *Segalovich I.* A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // *Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications.* — 2003. — P. 8.

30. *Snowball* [Электронный ресурс] // [сайт]. URL: <http://snowball.tartarus.org/> (дата обращения: 19.04.2011).

31. *Porter M. F.* An algorithm for suffix stripping // *Program: electronic library and information systems.* — 1993. — Vol. 14, no. 3. — Pp. 130–137.

32. Морфологический анализатор *myaso* [Электронный ресурс] // [сайт]. URL: <http://myaso.jetspade.ru/> (дата обращения: 22.04.2011).

33. Морфологический анализатор *rumorphy* [Электронный ресурс] // [сайт]. URL: <http://packages.python.org/rumorphy/> (дата обращения: 26.04.2011).

34. *Schmid H.* Probabilistic part-of-speech tagging using decision trees // *Proceedings of the International Conference on New Methods in Language Processing.* — 1994. — P. 9.

35. *TnT – Statistical Part-of-Speech Tagging* [Электронный ресурс] // [сайт]. URL: <http://www.coli.uni-saarland.de/~thorsten/tnt/> (дата обращения: 26.04.2011).

36. *Brants T.* TnT: a statistical part-of-speech tagger // *Proceedings of the 6th Conference on Applied Natural Language Processing.* — 2000. —

Рр. 224–231.

37. *Thomas D., Fowler C., Hunt A.* Programming Ruby 1.9 (3rd edition): The Pragmatic Programmers' Guide. — The Pragmatic Programmers, 2010.

38. Rubinius : Use Ruby™ [Электронный ресурс] // [сайт]. URL: <http://rubini.us> (дата обращения: 12.01.2011).

39. JSON [Электронный ресурс] // [сайт]. URL: <http://json.org/> (дата обращения: 27.04.2011).

40. redhat.com | Enterprise Linux-Open Source Application for Servers built on Linux [Электронный ресурс] // [сайт]. URL: <http://www.redhat.com/rhel/server/> (дата обращения: 07.04.2011).

41. HTML5 [Электронный ресурс] // A vocabulary and associated APIs for HTML and XHTML [сайт]. URL: <http://dev.w3.org/html5/spec/> (дата обращения: 27.04.2011).

42. Toward Text-to-Picture Synthesis / A.B. Goldberg, J. Rosin, X. Zhu, C.R. Dyer // *NIPS 2009 Mini-Symposia on Assistive Machine Learning for People with Disabilities*. — 2009.

43. *Mihalcea R., Leong C.* Toward communicating simple sentences using pictorial representations // *Machine Translation — Springer*. — 2008. — Vol. 22, no. 3. — Pp. 153–173.

44. A text-to-picture synthesis system for augmenting communication / X. Zhu, A.B. Goldberg, M. Eldawy et al. // *Proceedings of the National Conference of Artificial Intelligence*. — 2007. — Vol. 22, no. 2. — P. 1590.

45. *Yoshii M., Flaitz J.* Second Language Incidental Vocabulary Retention: The Effect of Text and Picture Annotation Types // *CALICO journal*. — 2002. — Vol. 20, no. 1. — Pp. 33–58.

46. *Manning C., Schütze H.* Foundations of statistical natural language processing. — MIT Press, 1999. — Vol. 59.

47. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition). — Springer Series in Statistics, 2008.

48. *Forsyth D., Ponce J.* Computer vision: a modern approach. — Prentice Hall Professional Technical Reference, 2002.

# Приложение А Функционально–структурные модели



Рисунок А.1 — Функционально-структурная модель «Извлечь ключевые фразы из текста», узел А-0: контекстная диаграмма

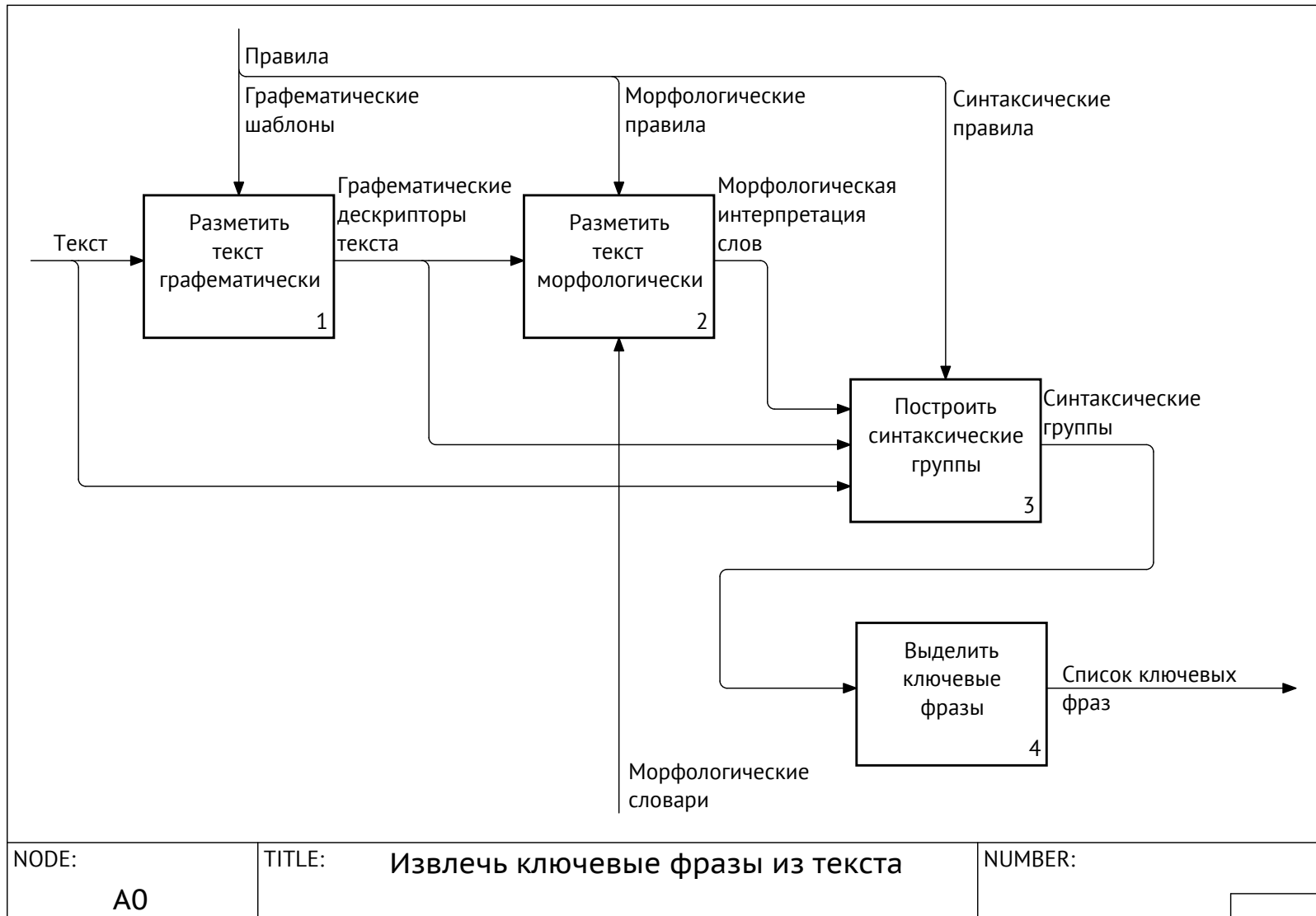


Рисунок А.2 — Функционально-структурная модель «Извлечь ключевые фразы из текста», узел А0: «Извлечь ключевые фразы из текста»

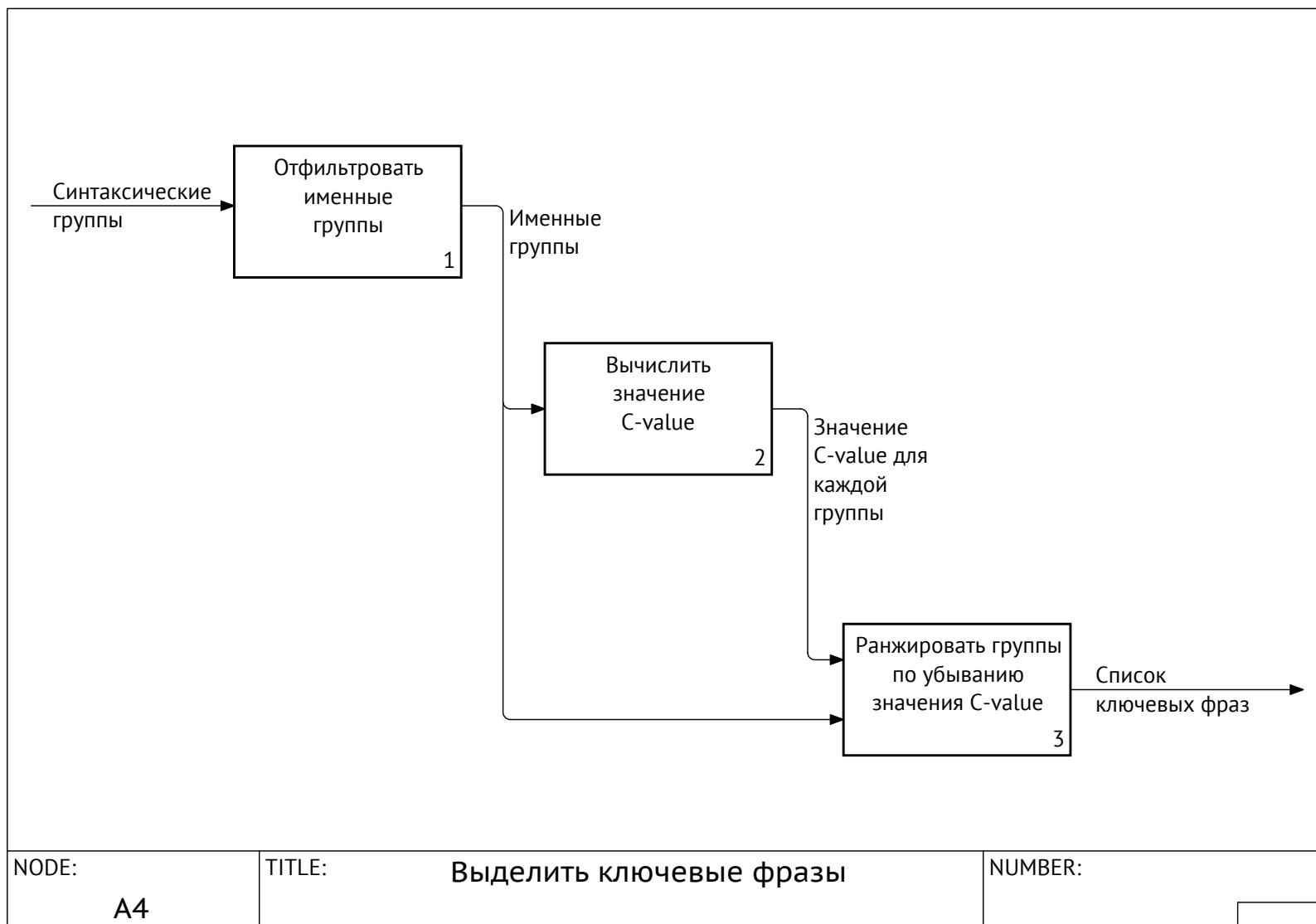


Рисунок А.3 — Функционально-структурная модель «Извлечь ключевые фразы из текста», узел А4: «Выделить ключевые фразы»

## Приложение Б Техническое задание

Министерство образования и науки Российской Федерации  
Федеральное государственное автономное образовательное учреждение  
высшего профессионального образования  
«Уральский федеральный университет имени первого Президента России Б. Н. Ельцина»

Физико–технологический институт

Кафедра вычислительной техники

УТВЕРЖДАЮ:

Зав. кафедрой, д. т. н., профессор

\_\_\_\_\_ С. Л. Гольдштейн

«\_\_\_\_\_» \_\_\_\_\_ 2011 г.

# СИСТЕМА АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ ФРАЗ ИЗ ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Техническое задание

на 16 листах

действует с «15» сентября 2010г.

Согласовано:

\_\_\_\_\_ А. Г. Кудрявцев

к. ф.-м. н., доц. каф. ВТ УрФУ

Екатеринбург

2011

## Содержание

1 Общие сведения . . . . .	61
1.1 Полное наименование разработки и её условное обозначение .	61
1.2 Шифр (номер) договора . . . . .	61
1.3 Наименование предприятий разработчика и заказчика системы	61
1.4 Перечень документов, на основании которых создаётся проект	61
1.5 Плановые сроки начала и окончания работы по созданию пакета	61
1.6 Порядок оформления и предъявления заказчику результатов работ по созданию системы . . . . .	62
2 Назначение и цели создания проекта . . . . .	62
2.1 Назначение и перечень объектов автоматизации . . . . .	62
2.2 Цели разработки . . . . .	62
2.3 Наименование и требуемые значения технических, экономических и социальных показателей, которые должны быть достигнуты .	63
2.3.1 Технический показатель . . . . .	63
2.3.2 Социальный показатель . . . . .	63
2.3.3 Экономический показатель . . . . .	63
3 Характеристика объектов автоматизации . . . . .	63
3.1 Краткие сведения об объекте автоматизации . . . . .	63
3.2 Сведения об условиях эксплуатации объекта и характери- стиках окружающей среды . . . . .	64
4 Требования к системе . . . . .	64
4.1 Требования к системе в целом . . . . .	64
4.1.1 Требования к защите информации от несанкциониро- ванного доступа . . . . .	64
4.1.2 Требования к численности и квалификации персонала	65
4.1.3 Требования к эргономике . . . . .	65
4.2 Требования к видам обеспечения . . . . .	65
4.2.1 Математическое обеспечение . . . . .	65
4.2.2 Информационное обеспечение . . . . .	65

4.2.3 Лингвистическое обеспечение . . . . .	65
4.2.4 Программное обеспечение . . . . .	65
4.2.5 Техническое обеспечение . . . . .	66
4.2.6 Организационное обеспечение . . . . .	67
4.2.7 Методическое обеспечение . . . . .	67
5 Состав и содержание работ по созданию работы . . . . .	67
5.1 Перечень стадий и этапов работ по созданию системы . . . . .	67
5.2 Сроки выполнения этапов работ . . . . .	68
6 Порядок контроля приёмки системы . . . . .	71
6.1 Общие требования к приёмке работ . . . . .	71
6.2 Виды испытаний . . . . .	72
6.2.1 Предварительные испытания . . . . .	72
6.2.2 Опытная эксплуатация . . . . .	72
6.2.3 Приёмочные испытания . . . . .	73
7 Требования к составу и содержанию работ по подготовке объекта автоматизации ко вводу Системы . . . . .	73
7.1 Требования к документированию . . . . .	73
Источники разработки . . . . .	74

## **1 Общие сведения**

### **1.1 Полное наименование разработки и её условное обозначение**

Тесуџк — система автоматического извлечения ключевых фраз из текста на естественном языке (далее — Система).

### **1.2 Шифр (номер) договора**

Выпускная квалификационная работа по специальности 230200.62 «Информационные системы».

### **1.3 Наименование предприятий разработчика и заказчика системы**

— Разработчик — Усталов Дмитрий Алексеевич, студент группы ФТ-47081 кафедры «Вычислительная техника» физико-технологического института ФГАОУ ВПО «УрФУ имени первого Президента России Б. Н. Ельцина».

— Заказчик — кафедра «Вычислительная техника» физико-технологического института ФГАОУ ВПО «УрФУ имени первого Президента России Б. Н. Ельцина».

### **1.4 Перечень документов, на основании которых создаётся проект**

Система создаётся на основе следующих документов:

- материалы спецпрактикума;
- учебный план подготовки по специальности 230200.62.

### **1.5 Плановые сроки начала и окончания работы по созданию пакета**

- начало работ: 15.09.2010г.
- окончание работ: 15.05.2010г.

## **1.6 Порядок оформления и предъявления заказчику результатов работ по созданию системы**

- в соответствии с ГОСТ 34.602 оформляется, согласуется и утверждается Техническое задание. Заказчику передаётся Техническое задание в виде отчёта;
- на основе Технического задания оформляется технорабочий проект;
- по окончании этапа проектирования оформляется руководство пользователя и разработчика;
- по окончании работ оформляется акт сдачи–приёмки выполненных работ;
- оформляется пояснительная записка к дипломной работе.

## **2 Назначение и цели создания проекта**

### **2.1 Назначение и перечень объектов автоматизации**

Назначение: автоматизация процесса выделения многословных терминов из текста на естественном языке.

Объект автоматизации: ручная обработка текста.

### **2.2 Цели разработки**

Глобальная цель — автоматизация процесса выделения многословных терминов из текста на естественном языке.

Локальные цели:

- упростить организацию поиска информационных ресурсов экспертом предметной области;
- ускорить процесс реферирования и агрегирования документов;
- обеспечить использование в качестве подсистемы интеллектуального анализа данных.

## **2.3 Наименование и требуемые значения технических, экономических и социальных показателей, которые должны быть достигнуты**

### **2.3.1 Технический показатель**

Сокращение временных затрат на выделение ключевых слов и фраз из текста на естественном языке в процессе индексации, аннотирования и реферирования документов. Использование системы должно заметно сокращать время, затрачиваемое экспертом предметной области при анализе документов: компьютер лучше справляется с обработкой больших объёмов информации.

### **2.3.2 Социальный показатель:**

- экономия рабочего времени персонала;
- снижение трудоёмкости получения результата путём замены ручного труда работой автоматической системы;
- повышение научного уровня за счёт возможности обработки больших корпусов текстов за небольшой отрезок времени;
- повышение квалификации персонала за счёт использования в его работе современных информационных технологий с простым и удобным пользовательским интерфейсом.

### **2.3.3 Экономический показатель**

Снижение себестоимости эксперимента достигается в результате:

- замены ручного труда работой автоматической системы;
- экономия временных ресурсов;
- экономия трудовых ресурсов;
- экономия энергетических ресурсов.

## **3 Характеристика объектов автоматизации**

### **3.1 Краткие сведения об объекте автоматизации**

Задача автоматизации — исследовать и реализовать систему автоматического извлечения ключевых фраз из текста на естественном языке таким

образом, чтобы она могла работать в качестве интеллектуальной подсистемы системы автоматической обработки документов.

Объект автоматизации — произвольный социоорганизационный.

### **3.2 Сведения об условиях эксплуатации объекта и характеристиках окружающей среды**

Оборудование, используемое при работе системы автоматического извлечения ключевых фраз из текста на естественном языке, находится в нормальных условиях окружающей среды. Серверное оборудование, содержащее хранилище данных, функционирует круглосуточно, ежедневно, с перерывами на обслуживание. Программное обеспечение, расположенное на рабочей станции пользователя, даёт возможность пользования системой ежедневно и круглосуточно.

## **4 Требования к системе**

### **4.1 Требования к системе в целом**

Система должна выполнять следующие функции:

- производить извлечение ключевых фраз из поступающих текстов на естественном языке;
- предоставлять пользователю подробную информацию о релевантности и значимости каждого потенциального термина в проанализированном тексте при помощи Web-интерфейса;
- обеспечивать расширяемость путём предоставления сотрудникам предприятия Заказчика собственного программного интерфейса.

#### **4.1.1 Требования к защите информации от несанкционированного доступа**

Система не требует принятия каких-либо мер по защите информации от несанкционированного доступа по причине отсутствия необходимости хранения в базе данных промежуточных вычислений, а также конечных результатов работы.

### **4.1.2 Требования к численности и квалификации персонала**

К работе с Системой должны допускаться сотрудники, имеющие навыки работы на персональном компьютере, ознакомленные с правилами эксплуатации и прошедшие обучение по работе с Системой.

### **4.1.3 Требования к эргономике**

Требования к эргономике заключаются в правильном построении пользовательского интерфейса системы. Интерфейс разрабатываемого пакета должен быть интуитивно понятным и доступным пользователю. Цветовая гамма интерфейса не должна оказывать утомляющего действия на пользователя.

## **4.2 Требования к видам обеспечения**

### **4.2.1 Математическое обеспечение**

Необходимо наличие функционально–структурных и алгоритмических моделей. Математическая модель должна быть подробной, для того чтобы можно было полностью осознать проблематику и методы решения поставленных задач.

### **4.2.2 Информационное обеспечение**

Интерфейс пакета должен быть интуитивно понятным и доступным пользователю, владеющему ПЭВМ на уровне пользователя. При необходимости пользователь обращается к руководству пользователя.

### **4.2.3 Лингвистическое обеспечение**

Пакет должен быть реализован на платформе Rubinius по принципу «тонкого клиента». Взаимодействие пользователя с системой осуществляется через Web-интерфейс.

### **4.2.4 Программное обеспечение**

При выборе программного обеспечения предпочтение должно отдаваться архитектурным решениям и программным продуктам, уже доказавшим свою

пригодность при решении подобных задач. Базовое программное обеспечение должно поддерживать и использовать стандартные сетевые протоколы передачи данных.

Для создания Системы используется следующее программное обеспечение:

- операционная система Fedora Linux;
- язык программирования Ruby: современный динамический объектный язык с элементами функционального программирования;
- программный каркас для разработки Web-приложений Sinatra;
- среда разработки: текстовый редактор Sublime Text 2 или аналогичный;
- система управления исходным кодом: Git или любая другая распределённая система контроля версий;
- система хранения пар «ключ–значение» Tokyo Cabinet;
- Web-сервер nginx;
- облачная PaaS-система Cloud Foundry.

Все перечисленные программные решения являются свободным программным обеспечением, отлично зарекомендовавшим себя при использовании в проектах различной сложности.

Пользователь должен иметь возможность воспользоваться Системой при помощи любого популярного Web-браузера (Internet Explorer, Mozilla Firefox, Google Chrome, Opera).

#### **4.2.5 Техническое обеспечение**

В состав комплекса технических средств должны входить:

- серверы баз данных;
- рабочие станции;
- сетевое оборудование;
- периферийное оборудование.

## **4.2.6 Организационное обеспечение**

Рекомендуется провести обучение пользователей работе с пакетом (демонстрация).

## **4.2.7 Методическое обеспечение**

Необходимо наличие документа – руководства пользователя по работе с системой автоматического извлечения ключевых фраз из текста на естественном языке.

# **5 Состав и содержание работ по созданию работы**

## **5.1 Перечень стадий и этапов работ по созданию системы**

Следует ориентироваться на основные стадии (системное проектирование, создание системно–обоснованного Технического задания, эскизный проект, технорабочий проект) и этапы (содержательное и концептуальное, полужформализованное и математическое моделирование; а также конструкторско–технологическое представление) проектирования.

Разработка системы включает в себя следующие этапы:

- определение и анализ требований: определяется нормативно–техническая документация, на основе которой создаётся система, принимаются во внимание пожелания заказчика по функциям, которые должна содержать разрабатываемая система;
- формирование спецификаций: на основе полученных данных о системе, условий её эксплуатации и рекомендаций по улучшению работы с ней создаётся системно–обоснованное Техническое задание;
- проектирование: разработка модели системы, определение компонентов и их взаимосвязи между собой, определения перечня функций, выполняемых каждым компонентом и системой в целом;
- кодирование: создание программного кода системы на основе выбранного программного, лингвистического, математического и информационного обеспечения;
- тестирование и верификация: проверка работоспособности системы;

— документирование: подготовка необходимой документации методического, технического и организационного характера по сдаче, вводу в эксплуатацию данной системы;

— сопровождение: проведение различных мероприятий для обеспечения работоспособности системы и её дальнейшему развитию.

## 5.2 Сроки выполнения этапов работ

Состав и содержание работ на всех стадиях жизненного цикла Системы представлены в таблицах Б.1, Б.2, Б.3, Б.4, Б.5, Б.6, Б.7, Б.8.

Таблица Б.1 — Формирование требований к Системе.

№	Этапы работы	Сроки выполнения
1	Обследование объекта	5 дней
2	Формирование требований пользователя к Системе	10 дней
3	Оформление отчёта о выполненной работе и заявки на разработку Системы	10 дней

Таблица Б.2 — Разработка концепции Системы.

№	Этапы работы	Сроки выполнения
1	Изучение объекта	5 дней
2	Проведение необходимых научно-исследовательских работ	15 дней
3	Разработка вариантов концепции Системы и выбор концепции Системы, удовлетворяющего требования пользователя	10 дней
4	Оформление отчёта о выполненной работе	3 дня

Таблица Б.3 — Техническое задание.

№	Этапы работы	Сроки выполнения
1	Разработка и утверждение Технического задания на создание Системы	4 дня

Таблица Б.4 — Эскизный проект.

№	Этапы работы	Сроки выполнения
1	Разработка предварительных проектных решений по Системе и её частям	30 дней
2	Разработка документации на Систему и её части	14 дней

Таблица Б.5 — Технический проект.

№	Этапы работы	Сроки выполнения
1	Разработка проектных решений по Системе и её частям	30 дней
2	Разработка документации на Систему и её части	3 дня
3	Разработка заданий на проектирование в смежных частях проекта объекта автоматизации	14 дней

Таблица Б.6 — Рабочая документация.

№	Этапы работы	Сроки выполнения
1	Разработка документации на Систему и её части	7 дней
2	Разработка или адаптация программы	60 дней

Таблица Б.7 — Ввод в эксплуатацию.

№	Этапы работы	Сроки выполнения
1	Подготовка объекта автоматизации ко вводу Системы в действие	2 дня
2	Подготовка персонала	7 дней
3	Пуско-наладочные работы	2 дня
4	Проведение предварительных испытаний	4 дня
5	Проведение опытной эксплуатации	3 дня
6	Проведение приёмочных испытаний	2 дня

Таблица Б.8 — Сопровождение Системы.

№	Этапы работы	Сроки выполнения
1	Выполнение работ в соответствии с гарантийными обязательствами	7 дней
2	Послегарантийное обслуживание	—

## **6 Порядок контроля приёмки системы**

### **6.1 Общие требования к приёмке работ**

Состав, объём и методы испытаний должны быть определены в документах «Программа и методика испытаний», являющимся неотъемлемой частью соответствующей документации технического и рабочего проекта.

Приёмку работ должна осуществлять сформированная Заказчиком рабочая группа из представителей Заказчика и Исполнителя.

Каждый выполненный Разработчиком этап принимается рабочей группой Заказчика. При наличии положительного результата, окончательный приём работы осуществляется Приёмочной комиссией.

Место проведения приёмочных испытаний должны быть согласованы Исполнителем с Заказчиком на этапе технического и рабочего проектирования. Сроки испытаний могут быть скорректированы Исполнителем и Заказчиком на этапе технического и рабочего проектирования.

Все виды испытаний должны отвечать требованиям ГОСТ 34.603-92.

По результатам своей работы Приёмочная комиссия оформляет Акт приёмки работ, который подписывается всеми членами Приёмочной комиссии и представляется на утверждение Заказчику.

## **6.2 Виды испытаний**

### **6.2.1 Предварительные испытания**

Предварительные автономные испытания специального программного обеспечения Системы проводятся в соответствии с программой и методикой предварительных автономных испытаний, подготовленной для каждого из функциональных модулей Системы с использованием автономных тестов, подготовленных Разработчиком и согласованных с Заказчиком.

Комплексный тест должен обеспечивать проверку выполнения функций специального программного обеспечения (функциональных модулей), установленных настоящим Техническим заданием, в том числе всех связей между ними, а также проверку реакции подсистемы на некорректную информацию и аварийные ситуации.

### **6.2.2 Опытная эксплуатация**

Опытная эксплуатация проводится в реальном режиме работы организации Заказчика. Все функции, автоматизация которых предусмотрена в Системе, должны выполняться с использованием средств автоматизации.

### **6.2.3 Приёмочные испытания**

Приёмочные испытания проводятся в соответствии с Программой и методикой приёмочных испытаний путём выполнения комплексных тестов, подготовленных Разработчиком и согласованных с Заказчиком.

## **7 Требования к составу и содержанию работ по подготовке объекта автоматизации ко вводу Системы**

Для создания условий функционирования объекта автоматизации, при которых гарантируется соответствие создаваемой информационной системой требованиям, содержащимся в настоящем техническом задании, и возможность эффективного использования Системы, в организации Заказчика на этапе работ «Подготовка объекта автоматизации к вводу системы в действие» должен быть проведён следующий комплекс мероприятий:

- установка и модернизация оборудования;
- обучение персонала.

### **7.1 Требования к документированию**

Документы, создаваемые в процессе работ по созданию информационной системы, должны разрабатываться в соответствии с ГОСТ 34.602-89 и РД 50-34.698-90.

Перечень документов, подлежащих разработке в рамках создания Системы:

- «Программа и методика испытаний», ГОСТ 34.602-89;
- «Руководство пользователя», РД 50-34.698-90.

## Источники разработки

- ГОСТ 34.003-90 Автоматизированные системы. Термины и определения. — Введ. 01.01.90. — М.: Издательство стандартов, 1989.
- ГОСТ 34.601-90 Автоматизированные системы. Стадии создания. — Введ. 01.01.90. — М.: Издательство стандартов, 1989.
- ГОСТ 34.602-89 Комплекс стандартов на автоматизированные системы. Техническое задание на создание автоматизированной системы. — Введ. 24.03.89. — М.: Издательство стандартов, 1989.
- ГОСТ 19.102-77 Единая система программной документации. — Введ. 01.01.90. — М.: Издательство стандартов, 1989.
- ГОСТ 34.603—92 Виды испытаний автоматизированных систем. — Введ. 01.01.93. — М.: Издательство стандартов, 1989.
- РД 50-34.698-90. Автоматизированные системы. Требования к содержанию документов.