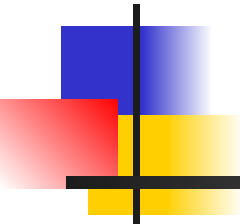


Дипломная работа

**ПРОГРАММНЫЕ СРЕДСТВА
ВЫЯВЛЕНИЯ
ТЕРМИНОЛОГИЧЕСКИХ
ВАРИАНТОВ В ТЕКСТАХ**



Антонов Вадим Юрьевич

Научный руководитель:
Ефремова Наталья Эрнестовна



ТЕРМИНЫ И ИХ ВАРИАНТЫ

- *Термины* – слова и словосочетания, называющие понятия предметной области
 - *рентгеновское излучение*
- Употребление терминов в текстах → *терминологические варианты*
 - *излучение, рентгеновские лучи*
- Выявление терминологических вариантов важно учитывать при построении тезаурусов, онтологий, предметных указателей, классификации текстов

КЛАССИФИКАЦИЯ ВАРИАНТОВ

Классификация терминологических вариантов для научно-технических текстов:

- < **графические** – компьютер/Компьютер
- < **флективные** – данные/данных
- < **орфографические** – браузер/броузер
- < **морфемные** – выполнение/исполнение
- < **сокращения** – высшее учебное заведение/ВУЗ
- < **синонимы** – абсорбция/поглощение
- < **лексико-синтаксические** –
центральный процессор/процессор,
текстовая коллекция/коллекция текстов



ПОСТАНОВКА ЗАДАЧИ

- Изучить классификацию терминологических вариантов и подходы к их выявлению
- На базе классификации разработать методы выявления терминологических вариантов в научно-технических текстах на русском языке
- На их основе реализовать программные средства
- Провести тестирование разработанных методов



ПОДХОДЫ К ВЫЯВЛЕНИЮ

- Символьный (статистический) подход
 - Термин и его варианты – символы
 - Вычисляется функция близости для термина и его варианта, для выбора порогового значения используется статистика
 - ❖ Не требуется лингвистическая информация и словари
 - ❖ Используется для орфографических и флективных вариантов
- Лингвистический подход
 - Термин и его варианты – словосочетания
 - Анализируется синтаксическая структура словосочетания, применяются правила образования вариантов и эвристики
 - ❖ Используется для лексико-синтаксических вариантов
 - ❖ Применён для английского и французского языков, для русского языка не изучен



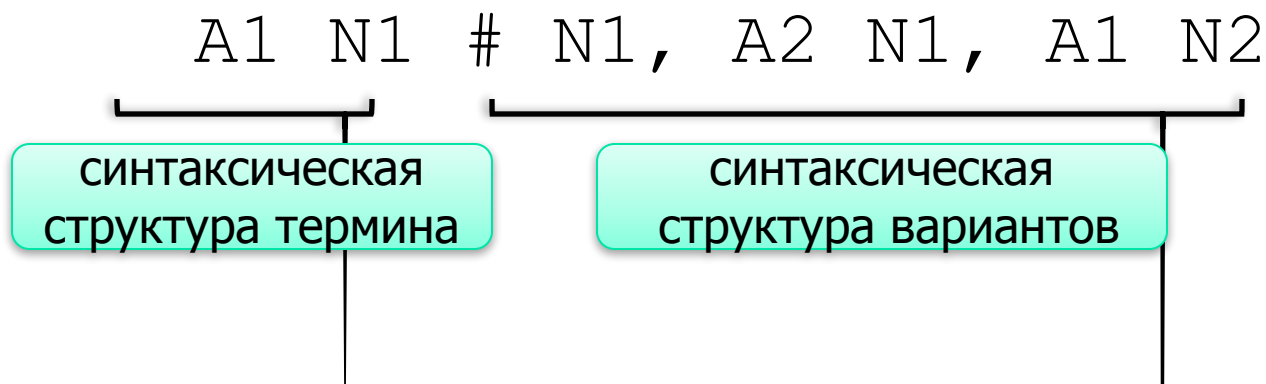
ПРЕДЛАГАЕМОЕ РЕШЕНИЕ

Для каждого типа терминологических вариантов используется свой метод, основанный на одном из подходов

Тип варианта	Метод выявления
Графические	приведение символов к одному регистру
Флективные	морфологический анализ
Орфографические	расстояние Левенштейна
Морфемные	словарь морфемного состава
Сокращения	эвристики для сокращений по первым буквам
Синонимы	словарь синонимов
Лексико-синтаксические	формальные правила образования вариантов

ЛЕКСИКО-СИНТАКСИЧЕСКИЕ ВАРИАНТЫ: ФОРМАЛИЗАЦИЯ

- Информация о лексико-синтаксических вариантах формализована в виде правил их образования
- Для формализации выбран язык LSPL и его библиотека:
 - позволяет описывать конструкции естественного языка в виде **лексико-синтаксических шаблонов**
 - предусмотрена возможность обработки информации, полученной в результате наложения LSPL-шаблона
- Правило образования – лексико-синтаксический шаблон вида:



ЛЕКСИКО-СИНТАКСИЧЕСКИЕ ВАРИАНТЫ: ВЫЯВЛЕНИЕ

Основано на

автоматической конкретизации шаблона правила

согласование

словарь синонимов

$A1 \ N1 \ \langle A1=N1 \rangle \ \# \ A1 \ N2 \ \langle \text{Syn}(N1, N2), A1=N2 \rangle$

рентгеновские лучи # *рентгеновское излучение*

термин

терминологический вариант

1. Распознавание термина заданной структуры

$A1 \Rightarrow$ *рентгеновские*, $N1 \Rightarrow$ *лучи*

2. Нормализация слов термина

рентгеновские \Rightarrow *рентгеновский*, *лучи* \Rightarrow *луч*

3. Построение шаблона возможного варианта

$A1 \langle \text{рентгеновский} \rangle \ N2 \ \langle \text{Syn}(\text{"луч"}, N2), A1=N2 \rangle$

4. Поиск варианта в тексте по конкретизированному шаблону

ОБЩАЯ ПРОЦЕДУРА ВЫЯВЛЕНИЯ

Исходная информация:

- Список терминов L_1
- Список кандидатов в терминологические варианты L_2

Алгоритм:

1. Для каждого $T_i \in L_1$ рассматриваем все $V_j \in L_2$
2. Для пары T_i и V_j проверяем, являются ли они вариантами, путём применения методов в определённом порядке

Результаты работы:

- Для каждого термина из L_1 – список терминологических вариантов из L_2

**Порядок
распознавания
типа варианта**

Графический

Флективный

Синонимы

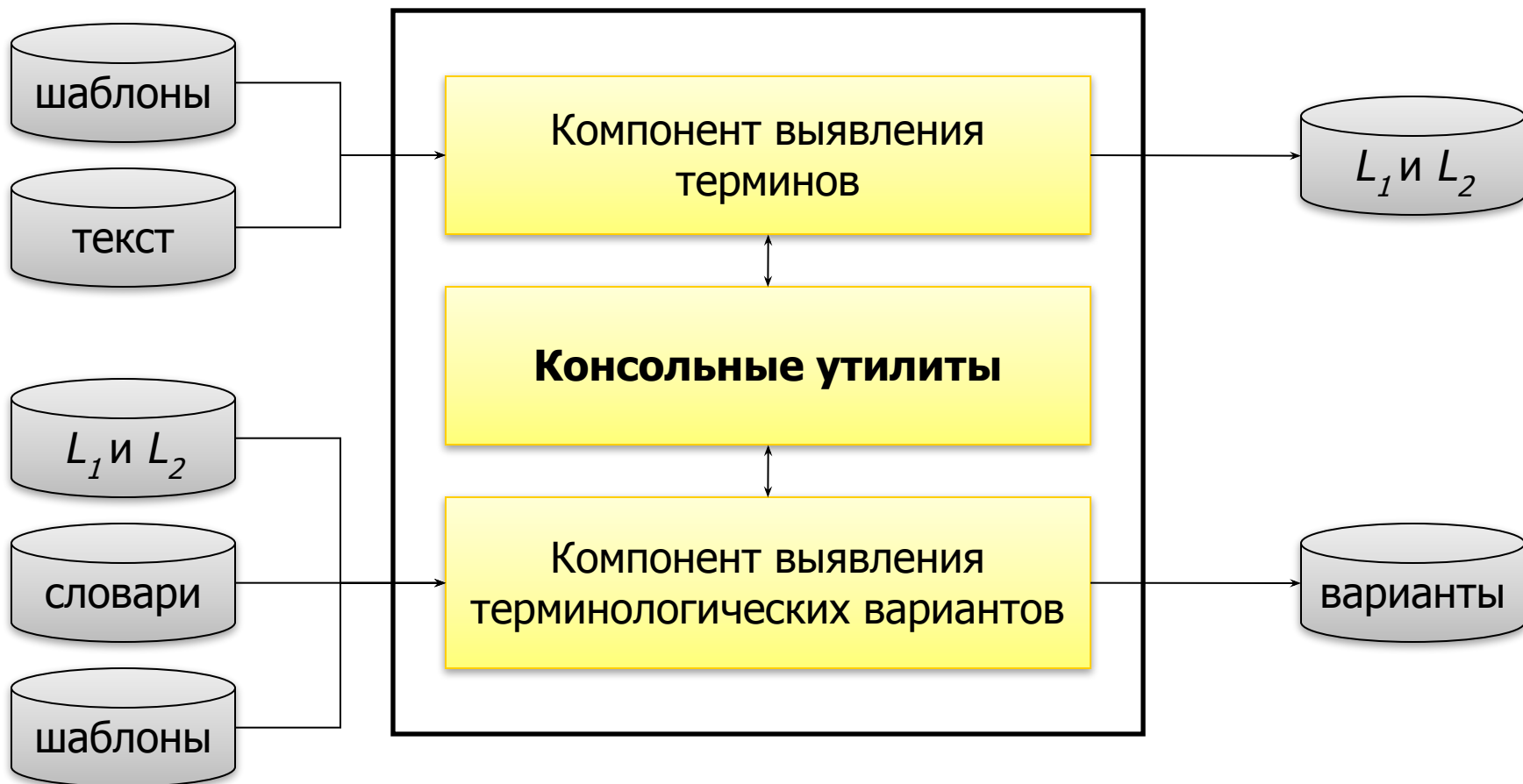
Морфемный

Лексико-
синтаксический

Сокращений

Орфографический

ПРОГРАММНЫЕ СРЕДСТВА: АРХИТЕКТУРА





РЕАЛИЗАЦИЯ И ТЕСТИРОВАНИЕ МЕТОДОВ

- Для реализации использован язык C++
- Библиотеки: LSPL, AOT, boost, STL
- Система контроля версий: git
- **Тестирование** на научно-технических текстах из областей физики и информатики объемом более 500кб
- Полнота выявления вариантов: 91%
- Точность выявления вариантов: 86%
 - Выявлено употреблений терминов без учета терминологических вариантов: 13668
 - Выявлено употреблений терминов с учетом терминологических вариантов: 25178
 - Процент прироста употреблений терминов: 84%



РЕЗУЛЬТАТЫ РАБОТЫ

- Проанализированы современные подходы к выявлению терминологических вариантов, изучена классификация вариантов, типичных для русскоязычных научно-технических текстов
- Разработаны методы выявления терминологических вариантов в соответствии с классификацией
- Библиотека языка LSPL расширена для формирования конкретизированных шаблонов
- Методы выявления реализованы в виде программных средств
- Тестирование показало состоятельность предложенных методов выявления

Результаты работы были представлены
(с публикацией) на:

- Международной научной конференции студентов, аспирантов и молодых ученых «Ломоносов 2010»
- Международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2010»

СПАСИБО ЗА ВНИМАНИЕ!



Расстояние Левенштейна

- Минимальное количество операций вставки, удаления и замены, необходимых для перевода одной строки в другую

=	=	ЗАМ	=	=	=	=
Б	Р	А	У	З	Е	Р
Б	Р	О	У	З	Е	Р

=	=	=	=	ЗАМ	ВСТ
М	А	С	С	А	
М	А	С	С	О	Й