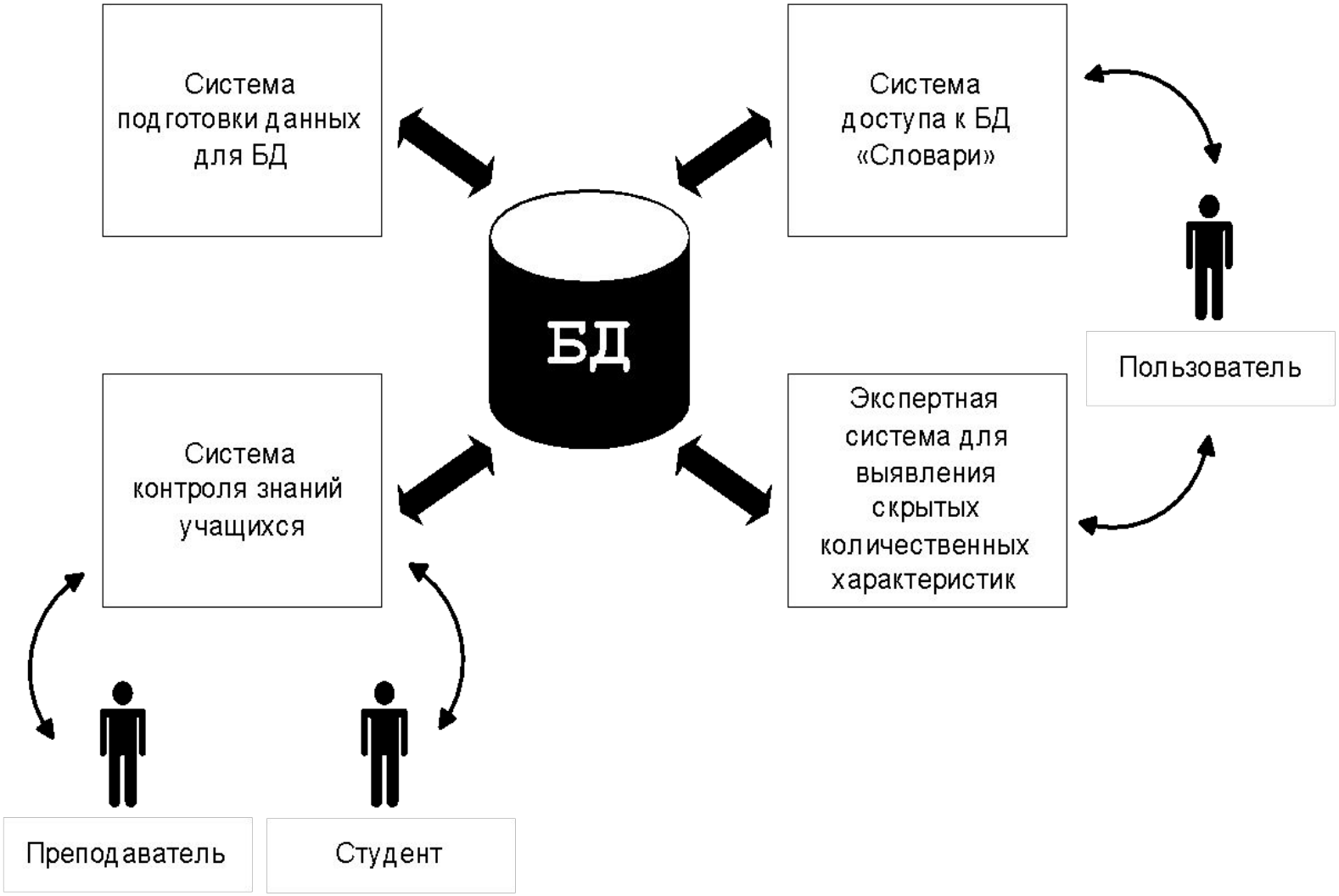
Петрозаводский государственный университет

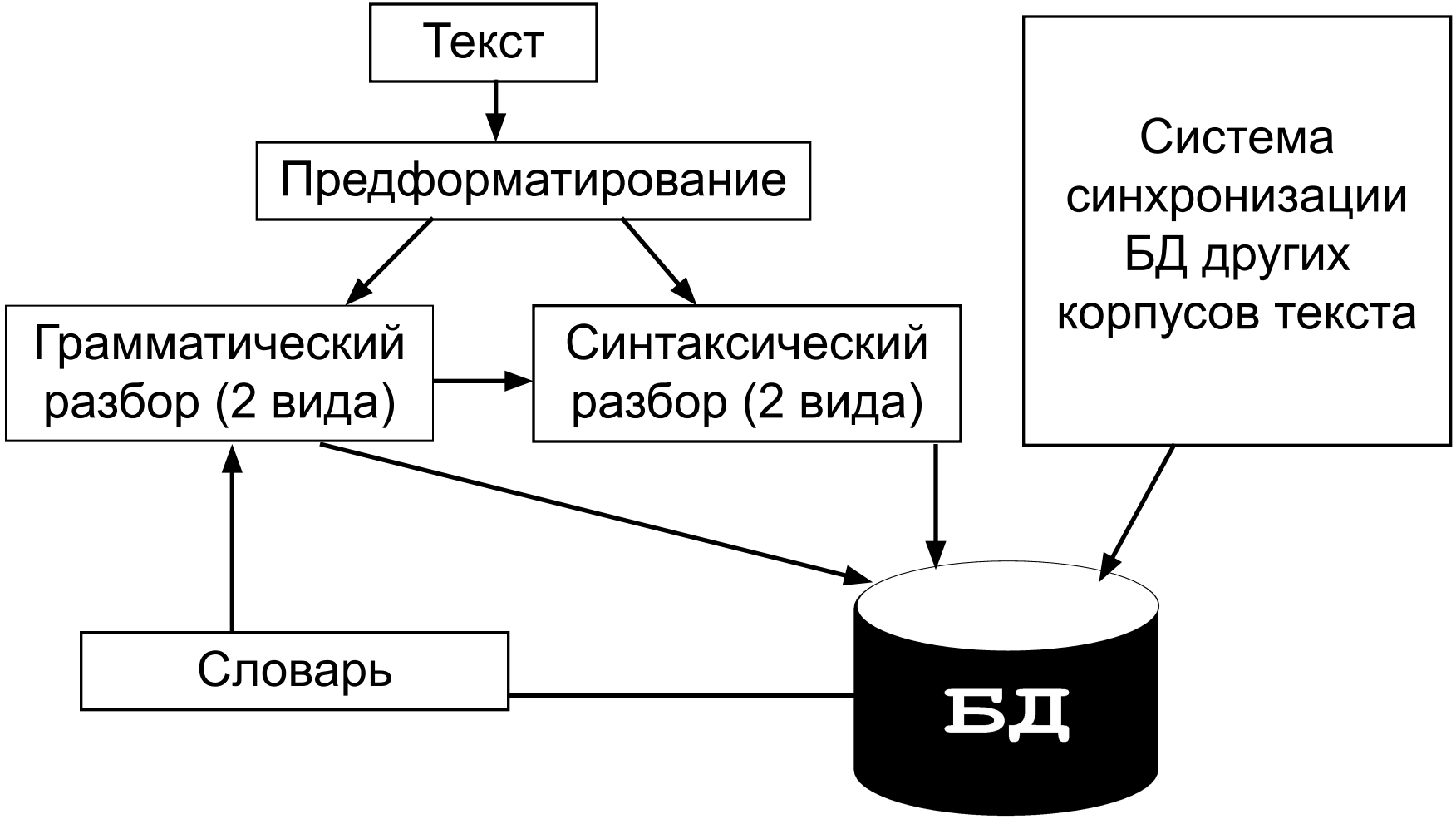
Программный комплекс «СМАЛТ». Морфологически размеченный корпус по русской публицистике второй половины XIX века

Авторы: ***Рогов А. А., Гурин Г.Б., Котов А.А., Сидоров Ю.В.,Суровцова Т.Г.***

Программный комплекс «СМАЛТ»

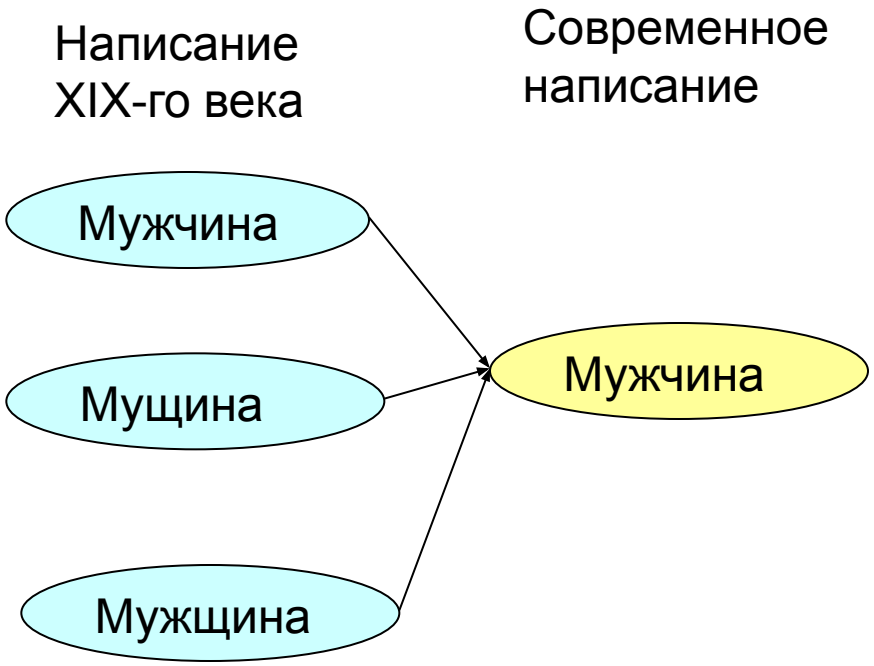


Система подготовки данных для БД



Особенности морфологического разбора текстов XIX века

(орфографическая вариативность)



Примеры: *очень-многіе, само-по-себе, до-сихъ-поръ, на-дняхъ, какъ-будто, ничемъ другъ-къ-другу необязанныхъ, студентскій миръ, взмахнутый, въ самомъ-достойнѣйшемъ, самоновѣйшій, низачто, само-малѣйшей, истинно-умные расположонъ, предстоитъ современем, состарѣлась, выростетъ, комунистъ, колосальный*

Виды морфологической разметки

Разметка 1

Опирается на следующий инвентарь частей речи: *существительное, прилагательное, числительное, местоимение, глагол, причастие, деепричастие, наречие, предикатив, союз, предлог, модально-дискурсивное слово или частица, междометие, компонент идиомы, антропоним*

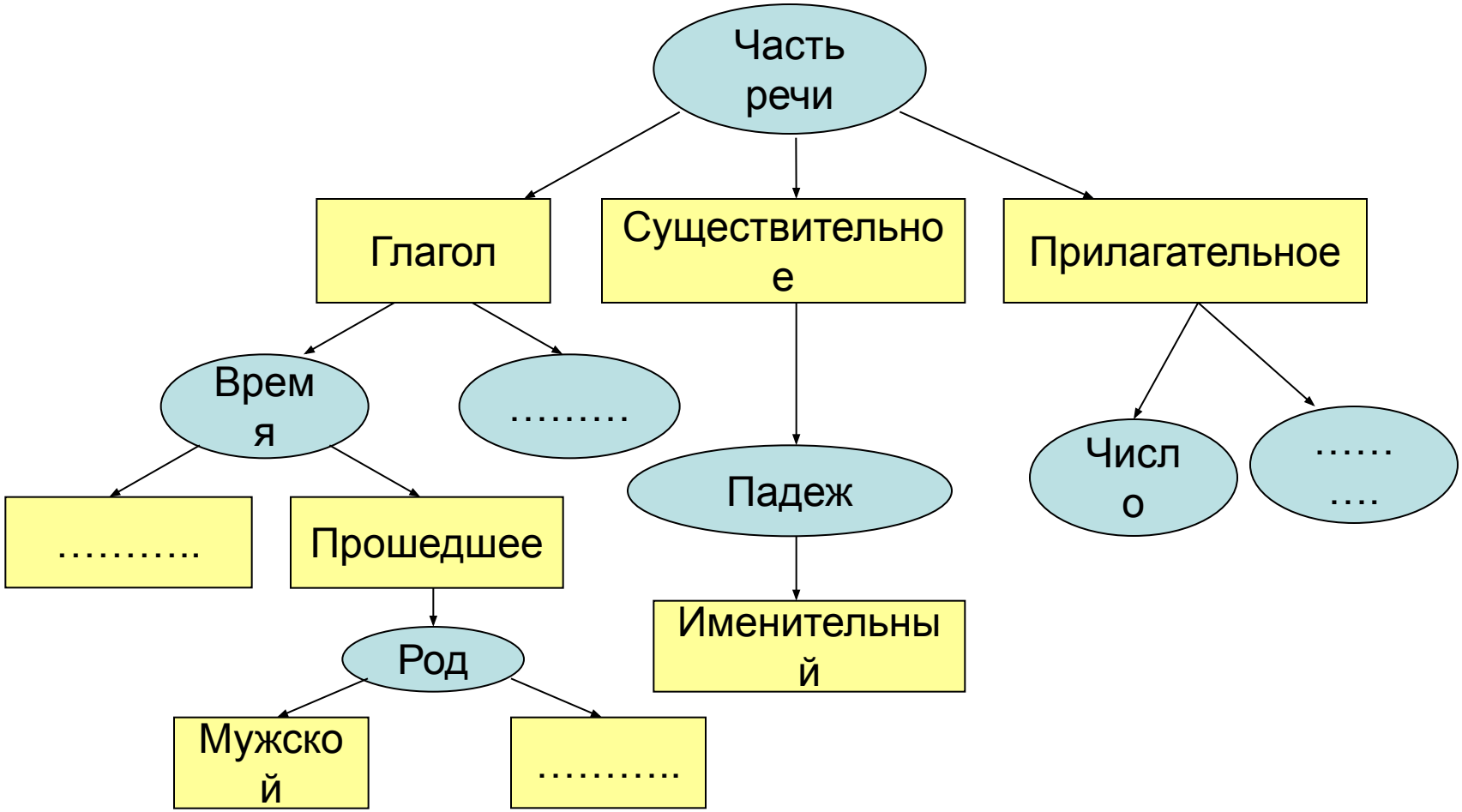
Предоставляет возможность поиска по значениям базовых морфологических категорий соответствующих частей речи.

Разметка 2

Ориентирована на школьную традицию и включает дополнительные грамматические параметры: лексико-грамматические разряды существительных. прилагательных, числительных, местоимений, типы склонения и спряжения.

Она предназначена для использования в образовательных целях и может рассматриваться как параллельный обучающий корпус.

Фрагмент структуры морфологической разметки



Статистика словаря

Корпус текстов состоит из публицистических статей разной тематической направленности из петербургских журналов «Время», «Эпоха», «Современник», «Гражданин» «Светоч», «Молва», «Библиотека для чтения», «Заря» XIX века в оригинальной орфографии.

	Первая разметка	Вторая разметка
Количество текстов	45	39
Количество слов в текстах	131435	113365
Количество словоформ в словаре	39901	30204
Общее число значений грамматических параметров	132	113

Преформатирование и грамматический разбор

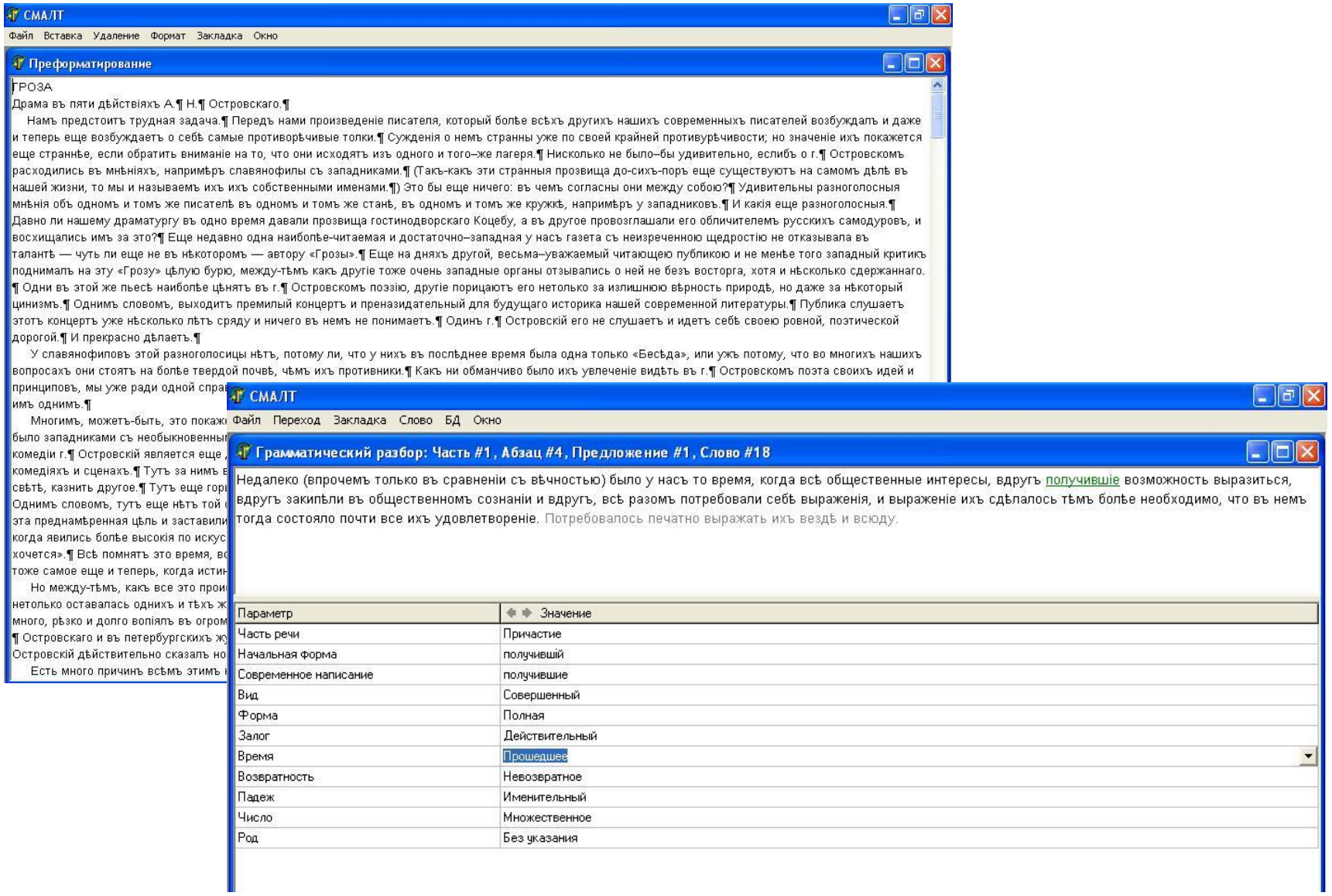


Схема доступа к БД

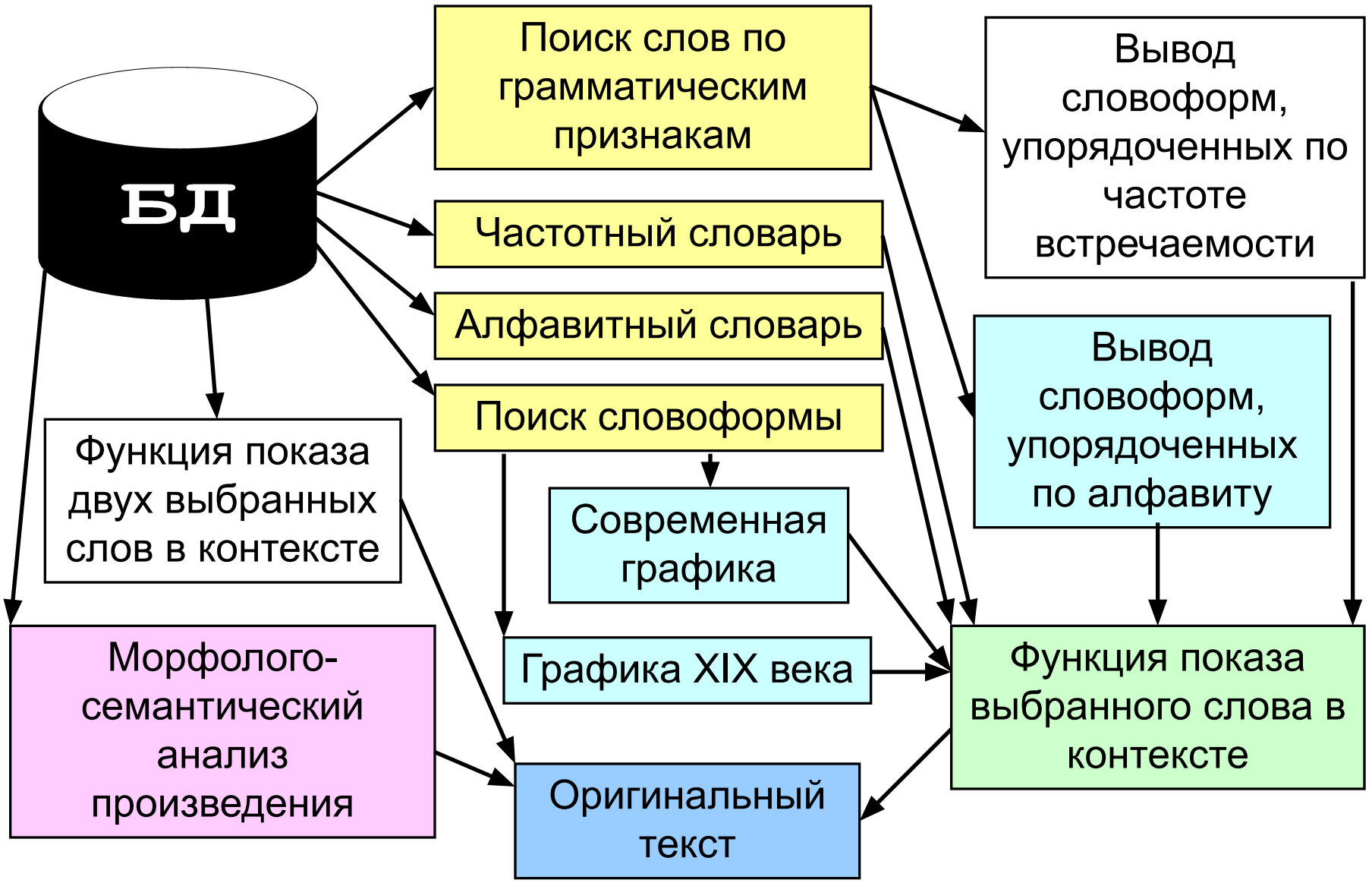
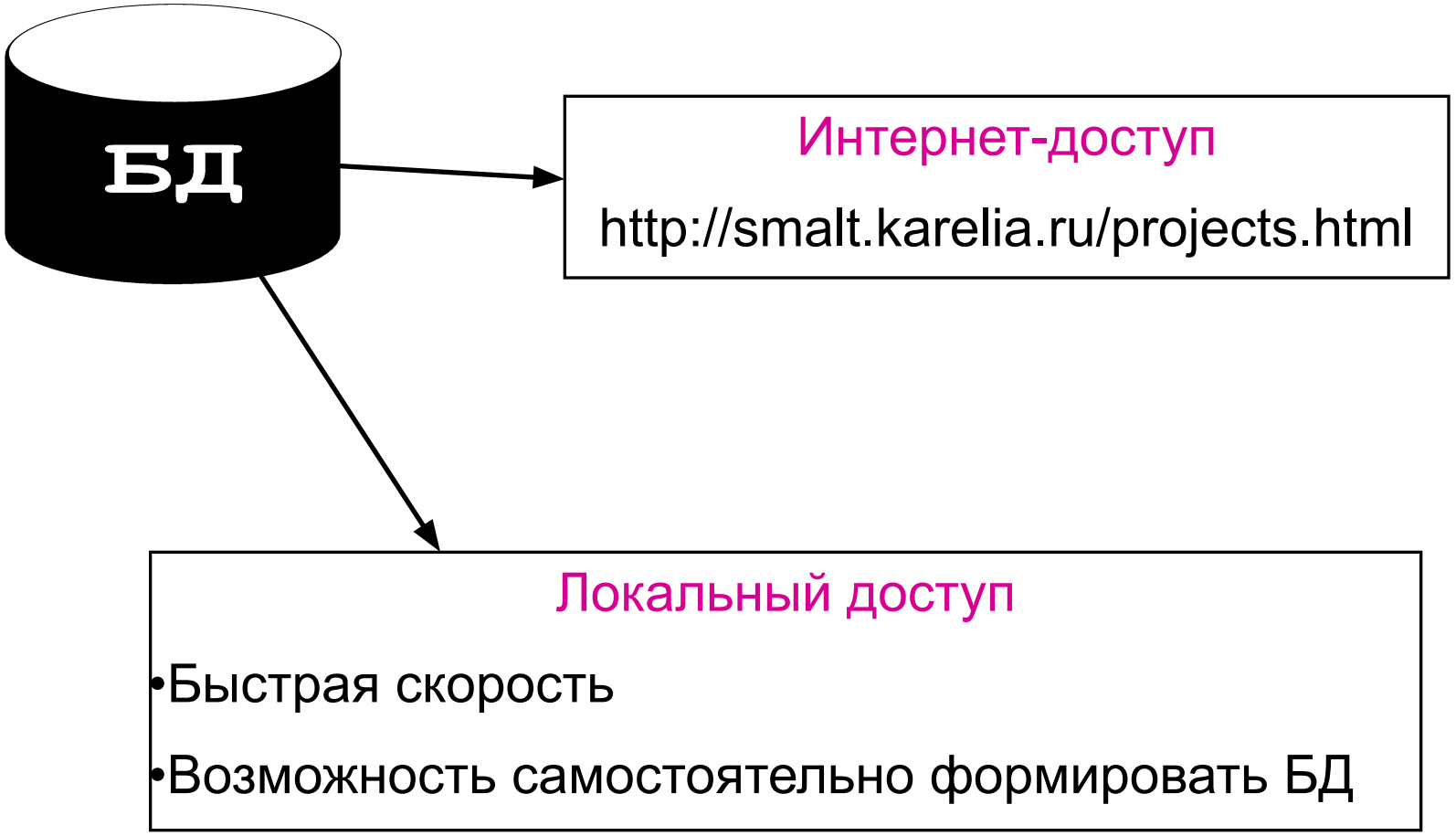
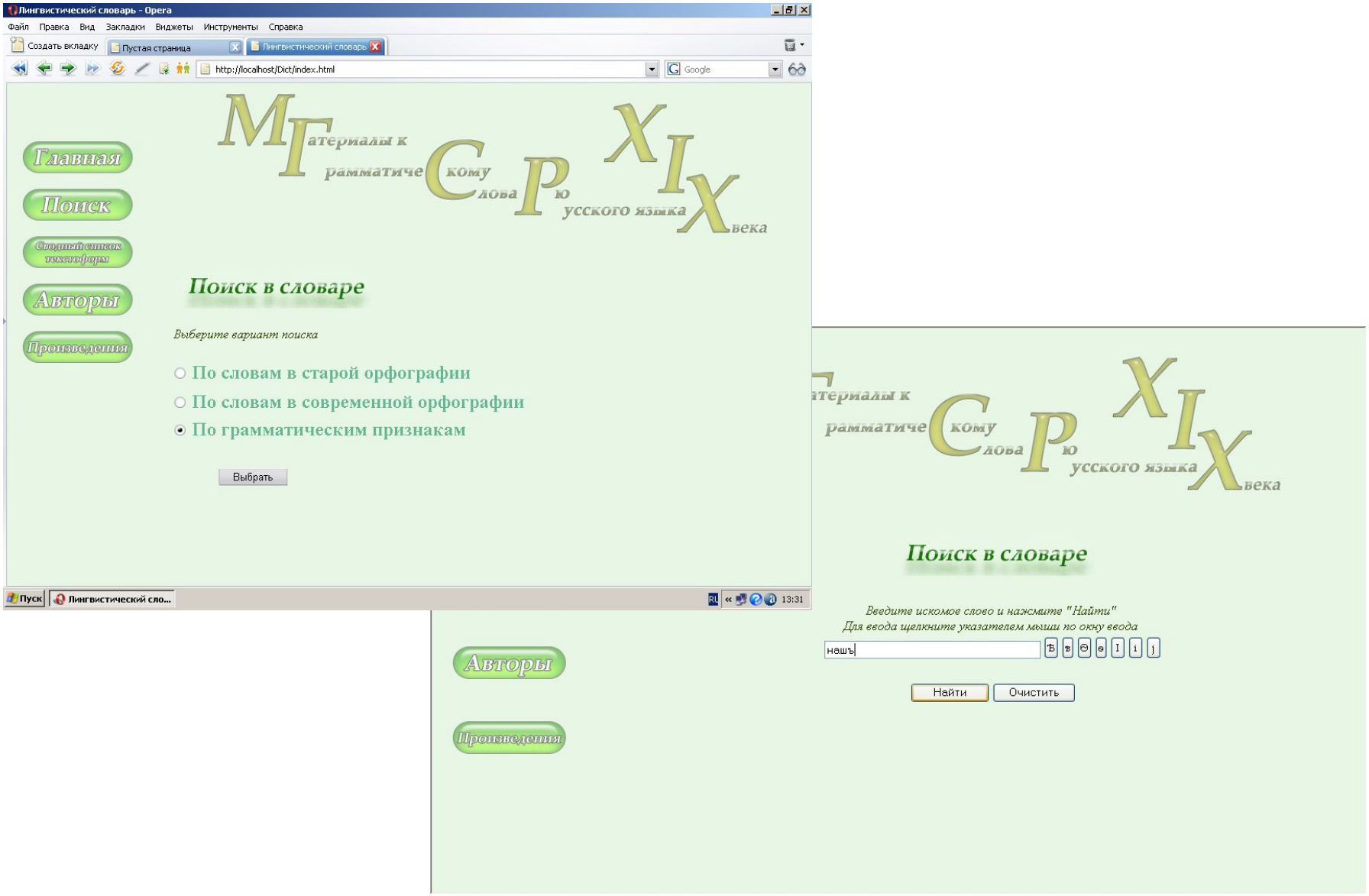


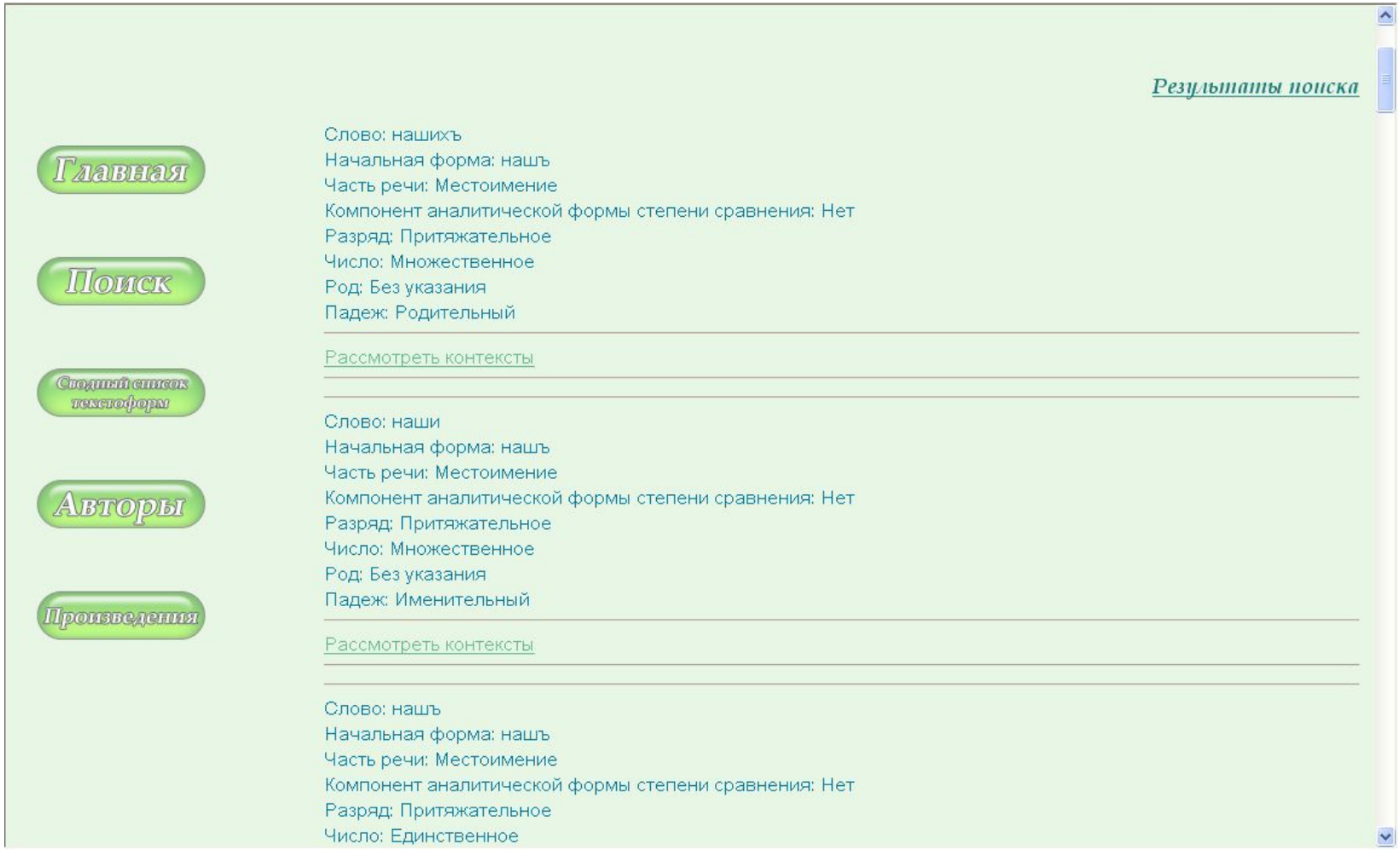
Схема доступа к БД



Поиск

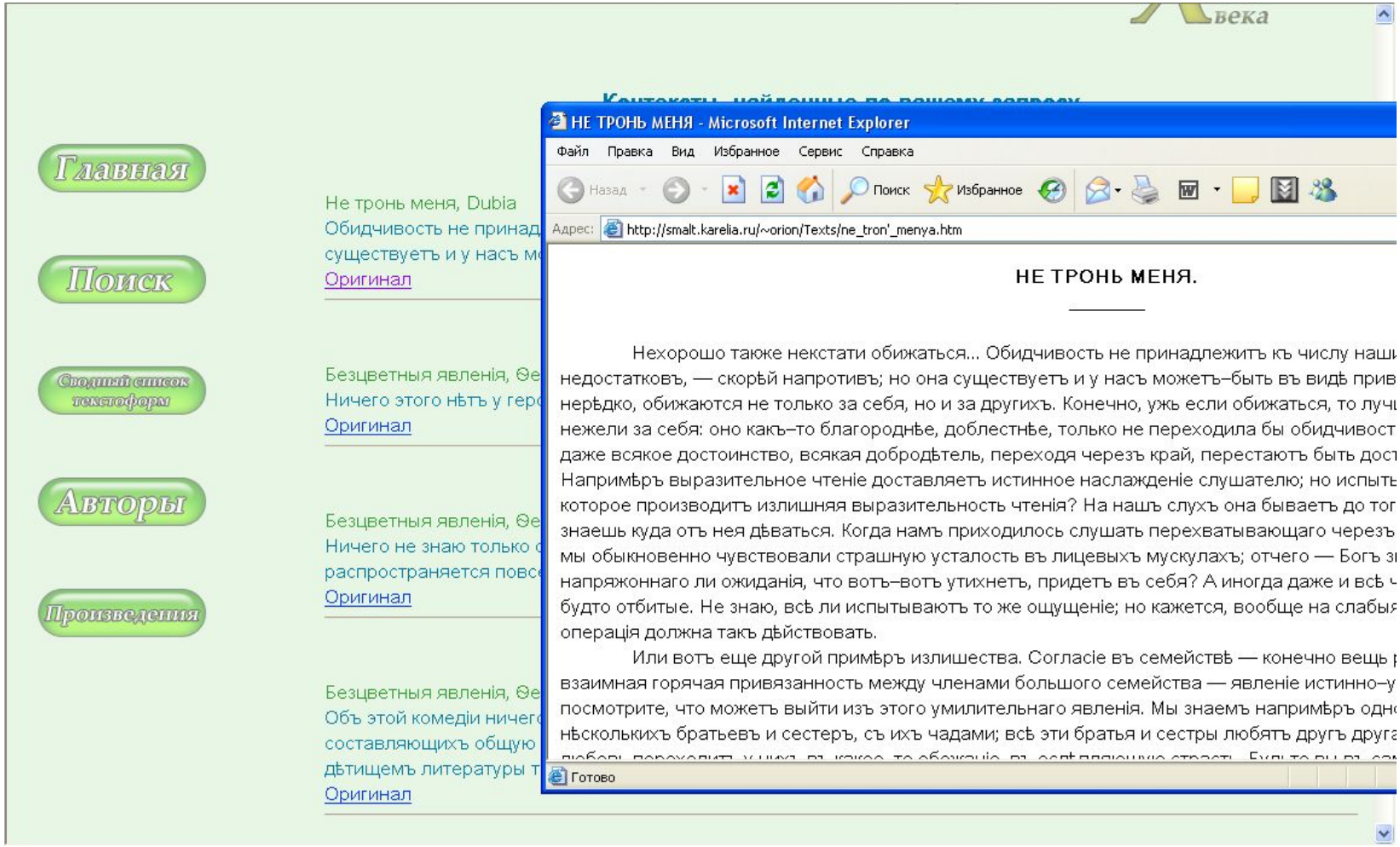

Лингвистический словарь - Opera
Файл Правка Вид Закладки Виджеты Инструменты Справка
Создать вкладку
Пустая страница
Лингвистический словарь
http://localhost/Dict/index.html
Google
Материалы к Грамматическому Словарю Русского языка XIX века
Главная
Поиск
Сводный список текстоформ
Авторы
Произведения
Поиск в словаре
Выберите вариант поиска
По словам в старой орфографии
По словам в современной орфографии
По грамматическим признакам
Выбрать
Пуск
Лингвистический сло...
13:31
Материалы к Грамматическому Словарю Русского языка XIX века
Поиск в словаре
Введите искомое слово и нажмите "Найти"
Для ввода щелкните указателем мыши по окну ввода
нашъ
Найти
Очистить
Авторы
Произведения

Вывод результата

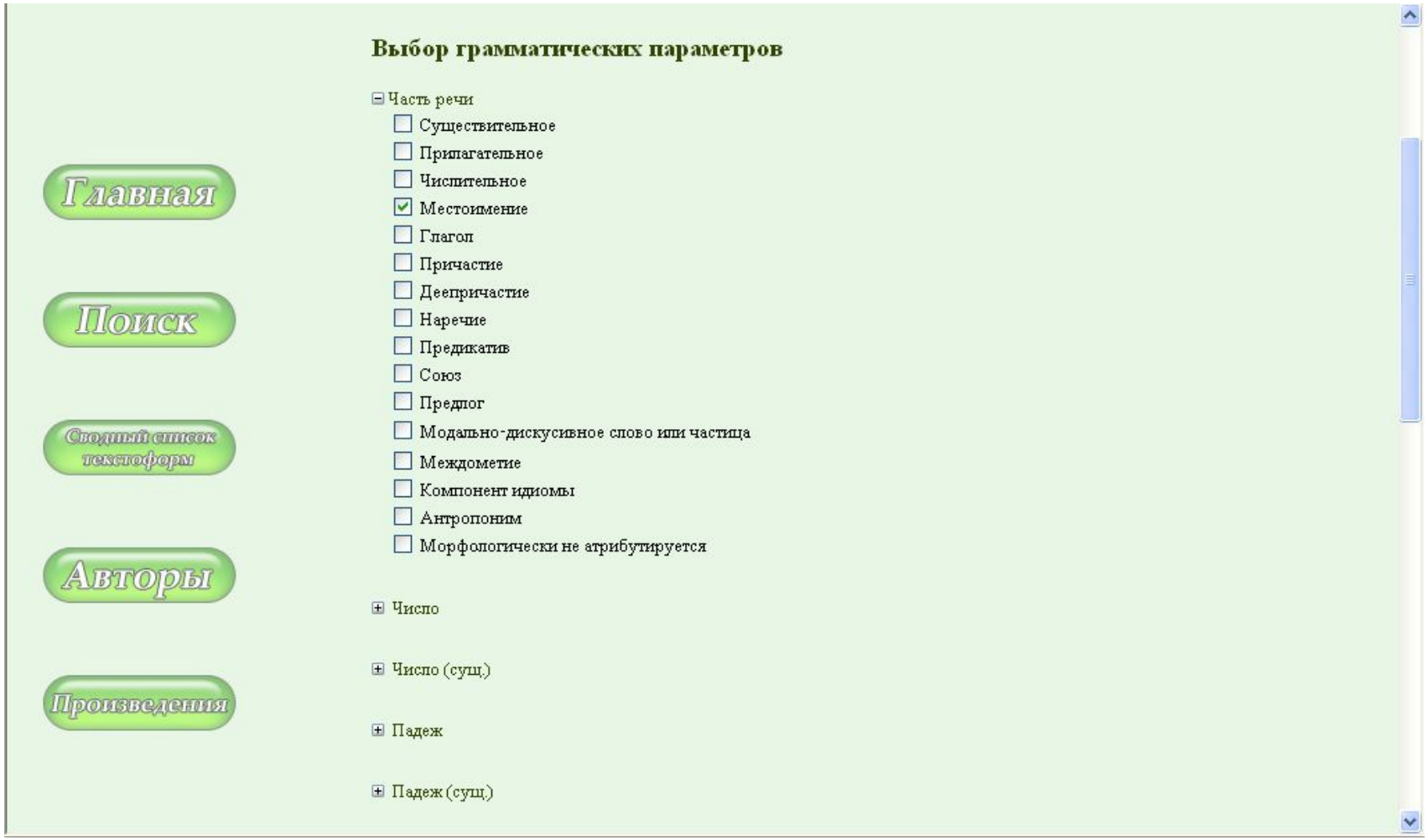


Вывод контекстов

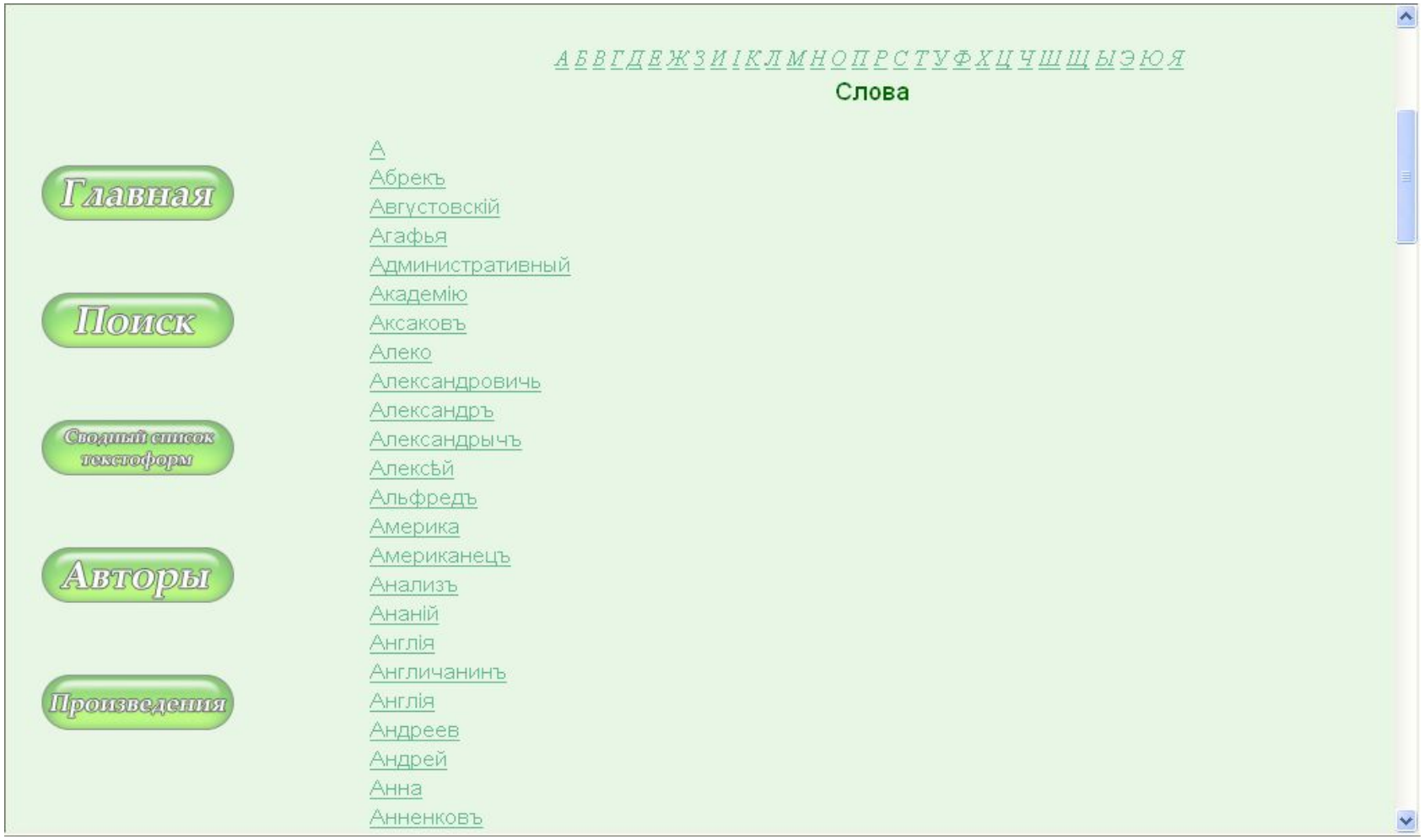
Вывод оригинального текста



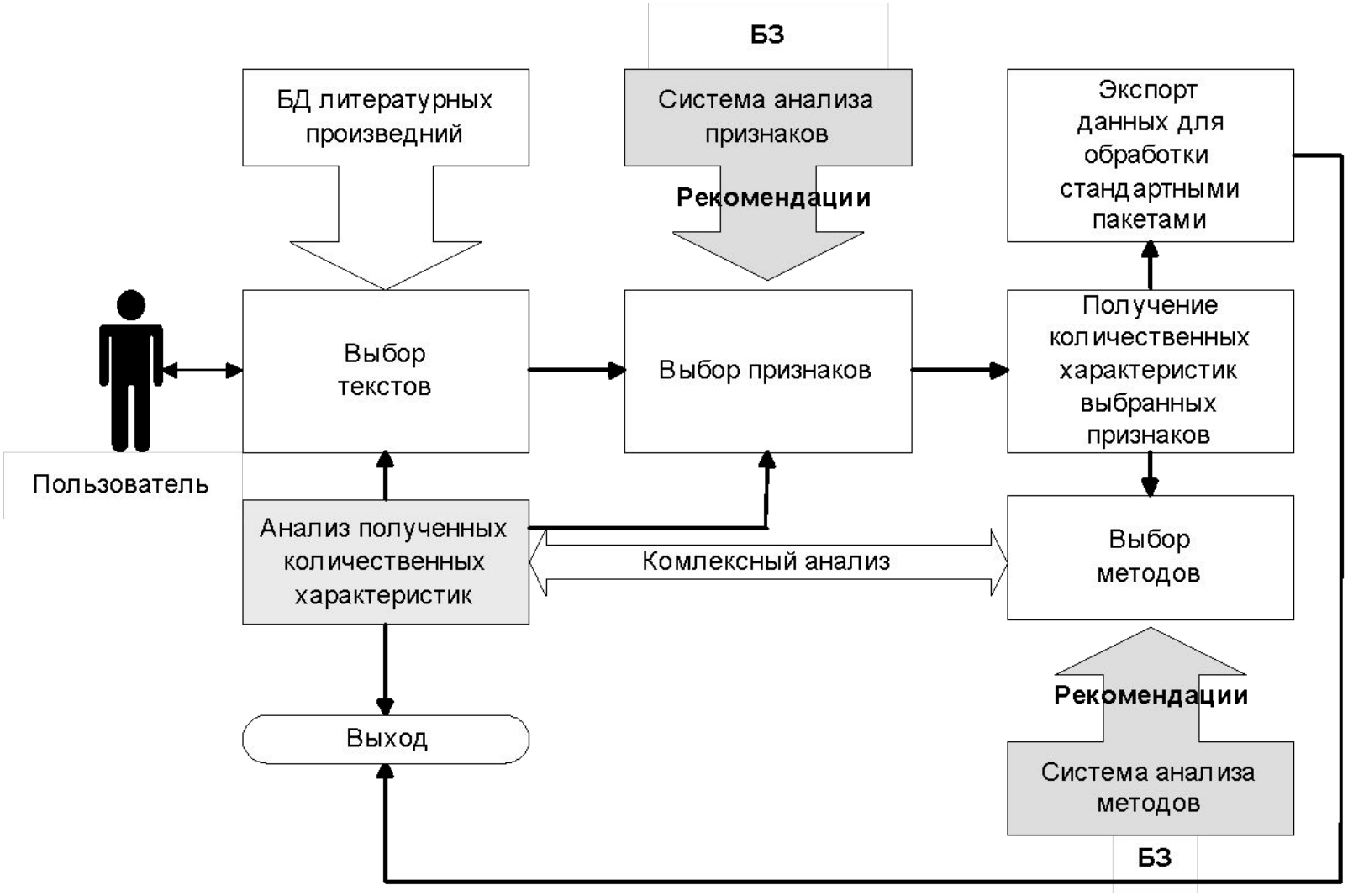
Поиск по грамматическим параметрам



Словарь


А Б В Г Д Е Ж З И І К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ы Э Ю Я
Слова
Главная
Поиск
Сводный список текстоформ
Авторы
Произведения
А
Абрекъ
Августовскій
Агафья
Административный
Академію
Аксаковъ
Алеко
Александровичь
Александръ
Александрычъ
Алексѣй
Альфредъ
Америка
Американецъ
Анализъ
Ананій
Англія
Англичанинъ
Англія
Андреев
Андрей
Анна
Анненковъ

Модуль статистического анализа программного комплекса «СМАЛТ»



Объект исследования

Достоевский редактировал и возглавлял три журнала

- Время (1861-1863)
- Эпоха (1864-1865)
- Гражданин (1873-1874)

Издавал свой личный журнал Дневник писателя (1876-1877, 1880-1881).

До сих пор остается открытым вопрос: какие же статьи из этих журналов действительно были написаны Ф.М. Достоевским?

Рабочим материалом исследования является 81 статья из Петербургских журналов «Время» и «Эпоха» (1861 – 1865 г.г.)

Методы анализа текстов

- **Статистические методы**
  - **Проверка статистических гипотез**
  - **Разбиение текстов на группы с использованием кластерного анализа**
- **Изучение переходов между составляющими единицами текста**
  - **Метод «сильного графа»**
  - **Метод подсчета отличий между матрицами**
- **Методы распознавания образов и искусственного интеллекта**
  - **Индуктивное построение статистических классификаторов**

Авторский инвариант

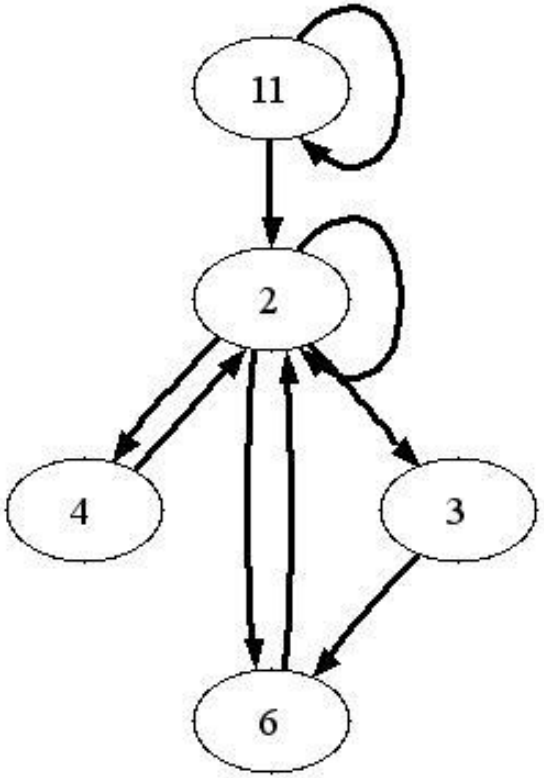
Под **авторским инвариантом** понимают такую характеристику литературных текстов (некий параметр), которая

1. однозначно характеризует своим поведением произведения одного автора или небольшого числа «близких авторов»,
2. принимает существенно разные значения для произведений разных групп авторов.

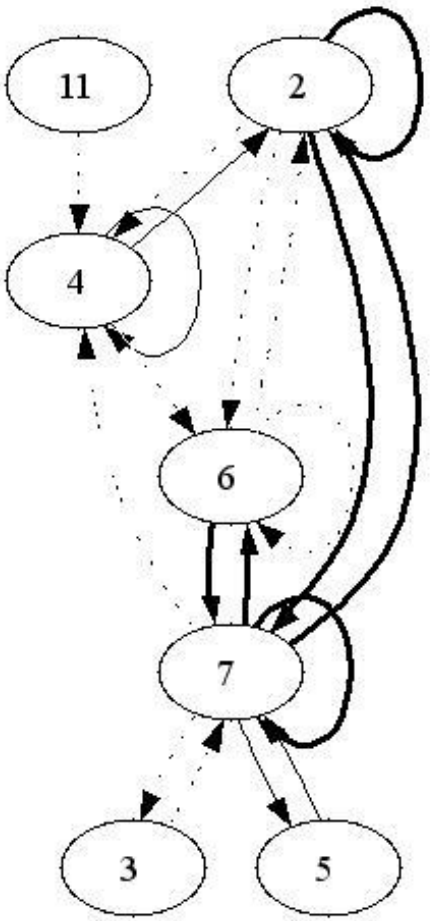
Желательно, чтобы число «разных групп» было достаточно велико, и чтобы каждая группа объединяла относительно мало похожих, близких по стилю авторов.

Свойства авторского инварианта

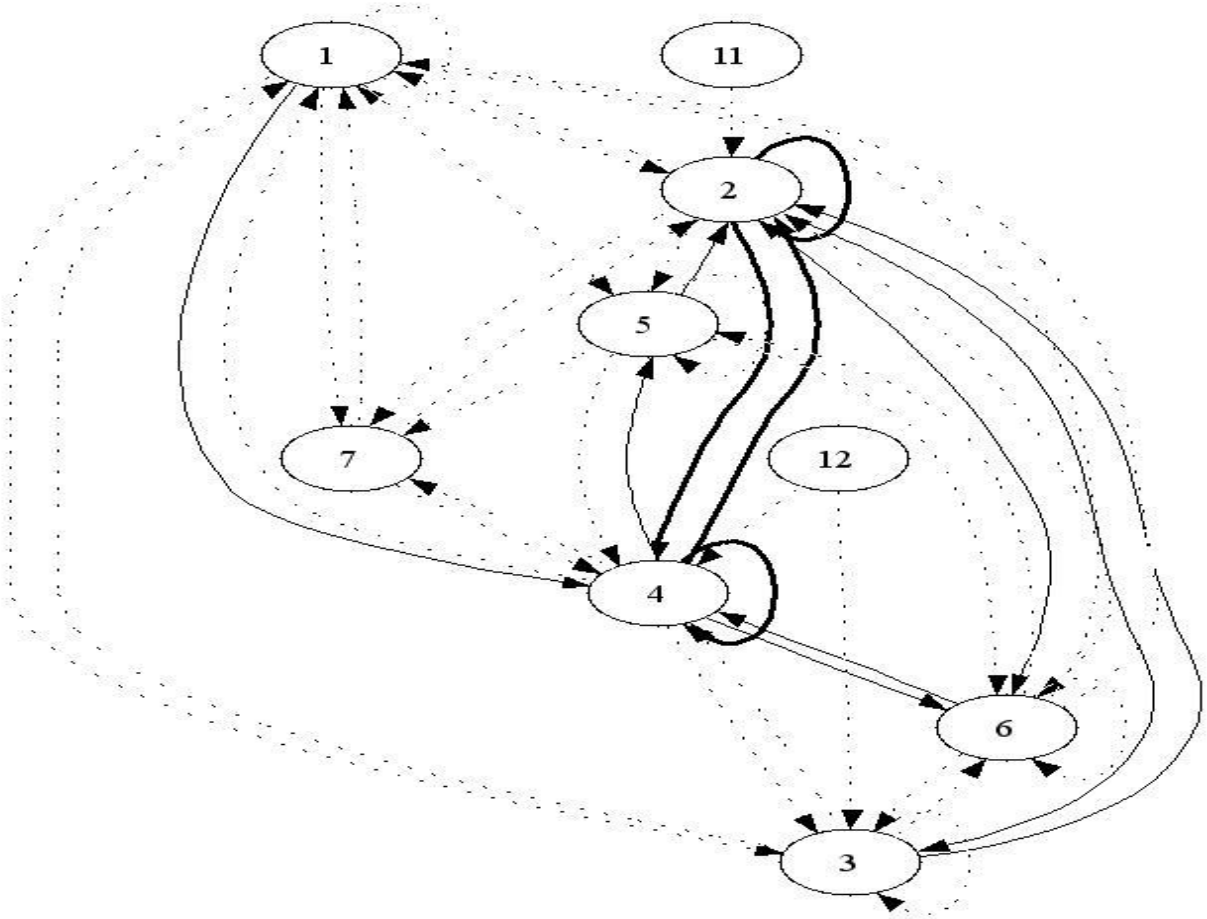
1. Она должна быть достаточно «массовой», интегральной, чтобы слабо контролироваться автором на сознательном уровне. Другими словами, она должна быть его «бессознательным параметром», коренящемся настолько глубоко, что автор даже не задумывается о нем.
2. Искомый параметр должен сохранять «постоянное значение» для произведений данного автора. То есть, иметь небольшое отклонение от среднего значения (слабо колебаться) на протяжении всех его книг.
3. Параметр должен уверенно различать между собой разные группы писателей. Другими словами, должно существовать достаточное число авторских групп, заметно отличающихся друг от друга значениями инварианта.



«Некролог. Иван Иванович Панаев», Dubia, «Время», 1862, №2, 16 предложений



«Несколько слов о Ристори», А. А. Григорьев, «Время», 1861, №2, 60 предложений



«Сильный граф» для произведения «Подписка на 1863 год», Ф. М. Достоевский, «Время», 1862, №9, 161 предложение

Часть таблицы коэффициентов близости текстов

Пороговое значение графа: **0,006** Узловое значение графа: **5**

	001	002	301	003	302	005	201	303	102	121	123	103	120	104	107
001		0,88	1	1	1	1	0,5	1	1	1	0,63	1	0,63	0,63	1
002			0,88	0,88	0,88	0,875	0,38	0,88	0,88	0,88	0,5	0,88	0,5	0,71	0,88
301				1	1	1	0,5	1	1	1	0,63	1	0,63	0,63	1
003					1	1	0,5	1	1	1	0,63	1	0,63	0,63	1
302						1	0,5	1	1	1	0,63	1	0,63	0,63	1
005							0,5	1	1	1	0,63	1	0,63	0,63	1
201								0,5	0,5	0,5	0,8	0,5	0,5	0,5	0,5
303									1	1	0,63	1	0,63	0,63	1
102										1	0,63	1	0,63	0,63	1
121											0,63	1	0,63	0,63	1
123												0,63	0,67	0,67	0,63
103													0,63	0,63	1
120														0,43	0,63
104															0,63
107															

Узловые значение графа устанавливались экспериментально и варьировались от 3 до 6.

Развитие исследования Гейра Хетсо

Хетсо Г. *Принадлежность Достоевскому: к вопросу об атрибуции Ф.М. Достоевскому анонимных статей в журналах "Время" и "Эпоха"*. SOLUM FORLAG A.S.: OSLO 1986.

Отличия:

1. Использование текстов в авторской орфографии и пунктуации;
2. Проверка устойчивости методик на разных объемах выборок;
3. Проверка гипотез о нормальности выборок, с целью правомерности использования некоторых статистических критериев;
4. Использование статей, автором которых Ф.М. Достоевский не является (например, статья А. Григорьева «Стихотворения А. С. Хомякова».)

Используемые лингвостатистические параметры

1. Средняя длина слова в буквах, вычисляемая на основании выборок размером в 200, 300, 400, 500 и 600 текстовых слов.
2. Общее распределение длины слова.
3. Средняя длина предложения в словах, вычисляемая на основании выборок размером в 30 предложений.
4. Общее распределение длины предложения.
5. Лексический спектр текста на уровне словаря.
6. Лексический спектр текста на уровне текста.
7. Индекс разнообразия лексики.

Средняя длина слова в буквах

$\mathbf{H_0}$ = {гипотеза о равенстве средних для двух выборок, одна из которых включает общую выборку по всем произведениям Достоевского}

Проверка данных выборки «Весь Достоевский» (ВД), состоящей из объединения 26 статей Достоевского на нормальность:

Объем выборки	Среднее	Дисперсия	Непараметрические критерии			Параметрический критерий		
			Колмогорова-Смирнова		Lilliefors	Хи-квадрат		
			Dn	p	p	X^2	Степени свободы	p
200	5,6190715	0,142409	0,0311863	n.s.	< 0,2	23,25523	16	0,1071657
300	5,6195259	0,1157226	0,0310931	n.s.	n.s.	10,51355	13	0,6514902
400	5,6216564	0,09697887	0,021458	n.s.	n.s.	3,864371	11	0,9737094
500	5,625463	0,08755674	0,0317953	n.s.	n.s.	10,71826	10	0,3799219
600	5,6199298	0,08039696	0,0311235	n.s.	n.s.	5,82722	9	0,757068

использовалась следующая формула критерия Стьюдента:

$$t = \frac{\overline{m}_1 - \overline{m}_2}{sd} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

В этой формуле m_1 и m_2 - сравниваемые средние частоты, n_1 и n_2- число выборок, и sd – несмещенная оценка среднего квадратичного отклонения в двух сериях выборок, вычисляемая по формуле:

$$sd = \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}}$$

Средняя длина слова в буквах. Критерий Стьюдента для разных объемов выборки.

№	Название статьи	200		300		400		500		600	
		t		t		t		t		t	
1	Введение	-1,33		-1,24		-1,17		-1,16		-1,00	
10	Ответ Русскому Вестнику	3,53		3,21		3,23	max	2,91	max	2,72	Max
26	Каламбуры в жизни и в литературе	-1,38		-1,42		-1,21		-1,18		-1,05	
27	Письмо постороннего критика в редакцию нашего журнала по поводу книг г-на Панаева и «Нового поэта»	-2,04		-1,85		-1,78		-1,54		-1,24	
30	Выставка в Академии художеств за 1860-1861 год.	6,56	>max	5,95	>max	5,64	>max	5,25	>max	5,04	>max
42	Стихотворения А.С. Хомякова	3,69	>max	3,11		3,16		2,94	>max	2,65	
47	Пожары	2,64		2,39		2,42		2,17		2,03	
48	Рассказы из народного русского быта Марка Вовчка	0,85		1,02		0,71		0,87		1,39	
49	Дурные признаки	6,94	>max	6,24	>max	5,87	>max	5,50	>max	5,28	>max
50	Гроза	-1,11		-1,05		-0,99		-1,05		-0,92	
51	Жуковский и романтизм	5,35	>max	4,69	>max	4,29	>max	4,10	>max	3,61	>max
52	Еще о петербургской литературе	0,61		0,43		0,38		0,38		0,41	
53	Пожары и зажигатели	3,69	>max	3,34	>max	2,81		2,45		2,83	>max
64	Тарас Шевченко	3,11		3,07		2,89		2,32		2,60	
65	Несколько слов о Писемском	3,93	>max	3,54	>max	3,17		2,96	>max	2,77	>max
66	Князь Серебряный	3,36		2,92		2,86		2,57		2,49	
80	Вместо фельетона	-3,24		-3,00		-2,80		-2,64		-2,55	
81	Внутренние новости	5,08	>max	4,59	>max	4,31	>max	4,02	>max	3,89	>max

Общее распределение длины слова.

Получены данные о том, сколько в каждом тексте слов, имеющих по 1, 2, 3, …16 и более буквам. Ставится вопрос: какова вероятность того, что распределения длин слов в буквах в двух статьях, одна и которых объединение статей Достоевского – ВД, взяты из одной и той же «генеральной совокупности» и могут рассматриваться как управляемые одними и теми же закономерностями?
непараметрический критерий Колмогорова-Смирнова, измеряющий разницу между накопленными частотами в сравниваемых текстах по формуле:

$$\lambda = d_{\max} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

где d_{max} обозначает максимальную разницу между накопленными относительными частотами, и n_1 и n_2 – количество слов в сопоставляемых текстах.

Средняя длина предложения в словах

Проводится тест исключительности на основании выборок в 30 предложений.

Проверка на нормальность дала положительный результат.

Подтверждается версия о том, что данный параметр обладает меньшей дискриминирующей силой, чем параметр «средняя длина слова».

Общее распределение длины предложения

Информация об общем распределении длины предложения была получена по интервалам в 1-5, 6-10, 11-15, …, 61 и более слов.

Использован непараметрический критерий Колмогорова-Смирнова.

Лексический спектр текста на уровне словаря
и
Лексический спектр текста на уровне текста

Лексический спектр текст - распределение частот слов в тексте.

Использовались частотные словари на каждые 500 слов текста. Все слова распределились в группы по 1, по 2, по 3, …, по 10 и более раз встречаемости в выборке. Далее определяем число слов в каждой группе, что означает распределение частот на уровне словаря, и «покрываемость» (Для определения лексического спектра на уровне текста рассматривается число словоформ в каждой группе, умноженное на частоту встречаемости слов из этой группы.) текста, что означает распределение частот на уровне текста. Если текст состоит из нескольких выборок, суммируются частоты встречаемости в тексте.

Индекс разнообразия лексики

Индекса разнообразия лексики - отношения числа разных слов к числу словоупотреблений.
Исследуется степень повторяемости в словаре писателя.

Общеизвестна тяга Достоевского к повторению одних и тех же слов и выражений.

Получены списки с указанием числа разных слов на каждые 200, 300, 400, 500 и 600 новых текстовых слов.

Проведен тест исключительности.

Результаты исследования

1. Несмотря на использование разных источников и соответственно на наличие некоторых различий, результаты с исследованием Хетсо совпали;
2. Показана правомерность использования статистического критерия Стьюдента с уровнем значимости *0,05* ;
3. Показана неустойчивость методик для некоторых параметров на разных объемах выборок;
4. Удалось показать неправомерность использования указанных параметров для атрибуции статей, являющихся материалом данного исследования.

Основной результат

В исследовании Хетсо был использован общий принцип применимости статистических методов. То есть для каждого метода определялась критическая граница $\alpha_{кр}$ и для каждой статьи определялся числовой параметр α. Далее делался вывод на основании двух гипотез: H_1 – {если $\alpha < \alpha_{кр}$, то статья скорее всего принадлежит Достоевскому}; H_2 – {если $\alpha > \alpha_{кр}$, то статья скорее всего не принадлежит Достоевскому}.

В данной работе удалось показать, что на данных методиках гипотеза H_1 не верна: в противном случае, следовало бы принять гипотезу о принадлежности Достоевскому статьи А. Григорьева «Стихотворения А. С. Хомякова», статей М. Достоевского «Рассказы из народного русского быта Марка Вовчка», «Пожары», «Гроза» и др.

Предположение о том, что распределение частей речи на первых трех и последних трех позициях предложения может быть авторским инвариантом

Уже целую неделю в Петербурге стояла ненастная погода.

Исследование проводилось:

1. Для каждого предложения текста
2. Для каждого первого и последнего предложения абзаца
3. Для каждого первого предложения абзаца
4. Для каждого последнего предложения абзаца

Исследование проводилось как с основным набором признаков (16), так и с расширенным (156):

1) с номера 1 по 7. Имя существительное (падеж)
2) с номера 8 по 13. Имя прилагательное (форма, степень сравнения)
3) с номера 14 по 25. Числительное (разряд по составу, разряд по значению)
4) с номера 26 по 34. Местоимение (разряд по значению)
5) с номера 35 по 45. Наречие (разряд по значению)
6) с номера 46 по 54. Наклонение (грамматическое значение, время)
7) с номера 55 по 102. Глагол (вид, залог, лицо)
8) с номера 103 по 106. Причастие, действительное, возвратное
9) с номера 107 по 112. Деепричастие, возвратное, одновременное с действительным глаголом – сказуемым
10) с номера 113 по 114. Модальное слово (синтаксические особенности)
11) с номера 115 по 117. Предлог (по составу)
12) с номера 118 по 126. Союз (по составу, по употреблению)
13) с номера 127 по 129. Частица (словообразующая функция)
14) с номера 130 по 147. Междометие (по образованию, по значению, синтаксические особенности)
15) с номера 148 по 154. Иностранное слово (язык)
16) с номера 155 по 156. Цитата (прозаическая/стихотворная).

Иерархическая кластеризация

Алгоритмы кластеризации:

- метод ближайшего соседа
- метод дальнего соседа

Меры близости между объектами:

1. Евклидова мера:

$$\rho(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

2. Мера Чебышева:

$$\rho(x, y) = \max_{i \in 1 \ldots n} |x_i - y_i|$$

В результате применения метода иерархической кластеризации оказалось, что невозможно четко выделить две группы объектов, ядро первой из которых преимущественно состояло бы из статей Достоевского, а ядро второй - из статей других авторов. Более того, во всех случаях на последних шагах объединения к основной группе присоединяются как атрибутируемые статьи, так и статьи Достоевского.

Одна из возможных причин – малый объемы текстов.

Необходимо отметить следующий факт, что при изучении деревьев иерархической кластеризации, можно заметить устойчивую тенденции к объединению в одну группу следующих объектов: 100, 202, 203. Под номером 100 обозначено объявление о подписке журнала "Время" с 1861 г., а под номерами 202 и 203 обозначены разные части объявления об издании журнала "Время" с 1861г. Естественно, что по содержанию и по стилистике эти тексты могли иметь много схожего, и быть написаны одним автором.

Оценка близости иерархических деревьев

Пусть n – число объектов объединения, тогда коэффициент близости 2 деревьев записывается как:

$$\rho = \frac{\sum_{k=1}^{n-1} \rho_k}{n-1}$$

где *n-1* показывает число уровней объединения или сечения, а

$$\rho_k = \frac{\sum_{i=1}^{n} \mu_i}{n}$$

где $\mu_i = \frac{2 \cdot N_i}{n_{i,1} + n_{i,2}}$ (1), либо $\mu_i = \frac{N_i}{n_{i,1} + n_{i,2} - N_i}$ (2)

где $n_{i,1}$ и $n_{i,2}$ – число объектов в группе, содержащей объект i, соответственно в первом и втором дереве, N_i – число совпадающих элементов в группах, содержащих объект i.

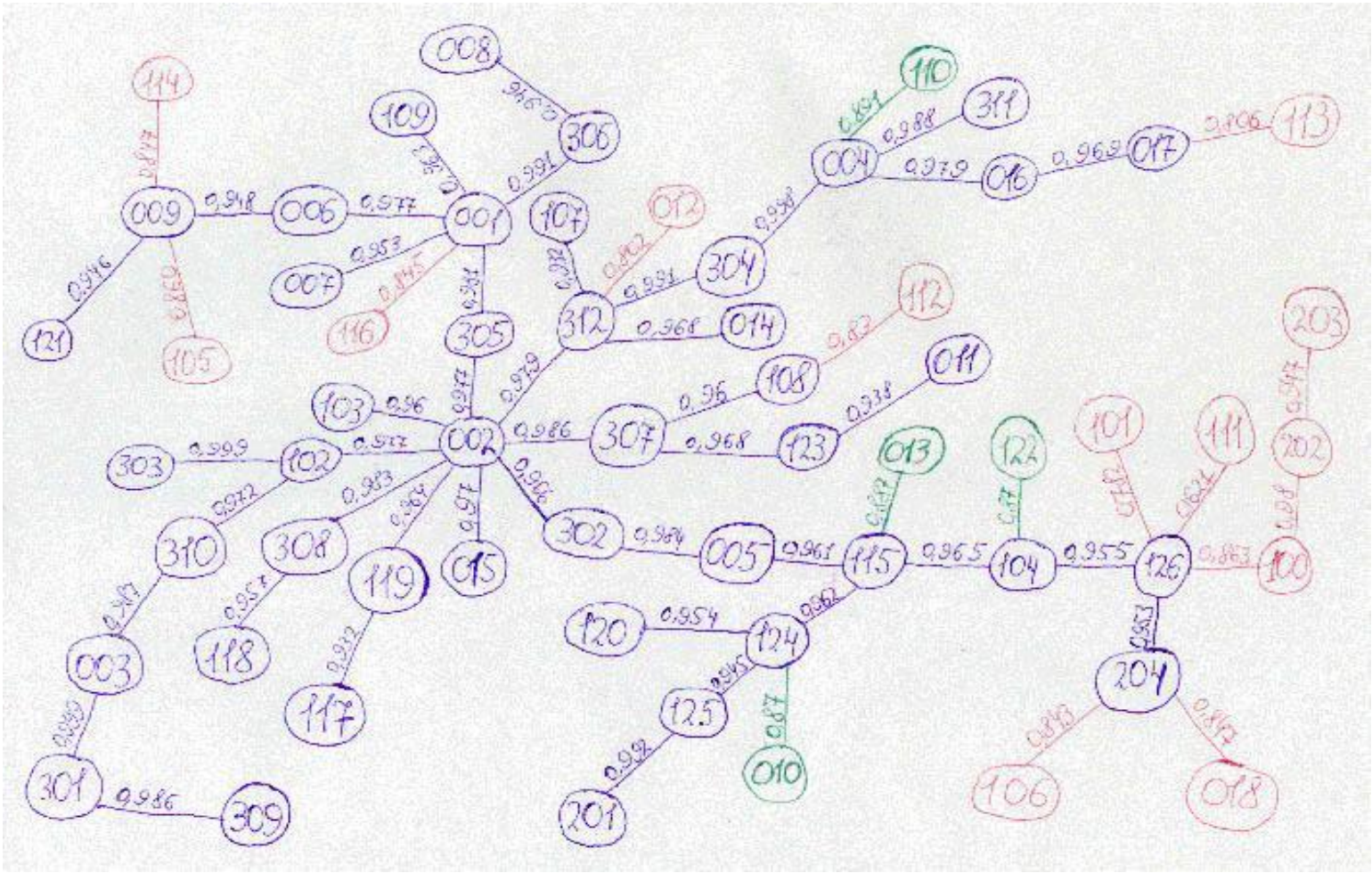
Оценка близости иерархических деревьев при соответствующих уровнях надежности и для разного числа объектов

n	0,05	0,1	0,2		0,8	0,9	0,95	0,99	0,999
4	0,639	0,639	0,667		0,840	0,861	1,000	1,000	1,000
5	0,592	0,600	0,615		0,718	0,798	0,821	0,917	1,000
6	0,567	0,576	0,587		0,681	0,725	0,761	0,847	0,932
59	0,402	0,405	0,409		0,424	0,428	0,431	0,440	0,450
60	0,402	0,404	0,408		0,423	0,427	0,431	0,438	0,447
61	0,401	0,404	0,407		0,422	0,426	0,430	0,436	0,447
62	0,401	0,404	0,407		0,422	0,426	0,429	0,437	0,449
63	0,401	0,403	0,407		0,421	0,425	0,429	0,437	0,443
98	0,389	0,391	0,393		0,403	0,406	0,409	0,414	0,419
99	0,388	0,390	0,393		0,403	0,406	0,409	0,413	0,419
100	0,388	0,390	0,392		0,403	0,406	0,408	0,413	0,417

Результаты

		Ближайший сосед		Дальний сосед	
		Евклидова мера	Мера Чебышева	Евклидова мера	Мера Чебышева
По всему тексту	(1)	0,877995	0,86083	0,747402	0,710619
	(2)	0,839487	0,81888	0,684595	0,630516
По первому и последнему предложению абзаца	(1)	0,936056	0,89733	0,839434	0,750736
	(2)	0,914381	0,86728	0,787565	0,683991
По первому предложению абзаца	(1)	0,910971	0,88638	0,840594	0,730016
	(2)	0,880166	0,85109	0,788498	0,648849
По последнему предложению абзаца	(1)	0,941005	0,91985	0,86204	0,768857
	(2)	0,921237	0,89372	0,813811	0,699331

Метод корреляционных плеяд



Пороговое значение	0.9	0.87
Выбывшие из основной группы	010, 012, 013, 018, 100, 101, 105, 106, 110, 112, 113, 114, 116, 122, 202, 203	012, 018, 100, 101, 105, 106, 112, 113, 114, 116, 202, 203