На правах рукописи

УДК 81:322; 81:372.88; 519.769

САЛОМАТИНА Наталья Васильевна

МЕТОДЫ И ПРОГРАММНЫЕ СРЕДСТВА ВЫДЕЛЕНИЯ И ЧИСЛЕННОГО ОЦЕНИВАНИЯ ВАРИАТИВНОСТИ ЯЗЫКОВЫХ ЕДИНИЦ

05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Новосибирск – 2009

Работа выполнена в Институте математики им. С.Л. Соболева СО РАН

Научный руководитель: Гусев Владимир Дмитриевич,
кандидат технических наук,
старший научный сотрудник

Официальные оппоненты: Хабаров Валерий Иванович,
доктор технических наук,
профессор

Сидорова Елена Анатольевна,
кандидат физико-математических наук

Ведущая организация: Научно-исследовательский
вычислительный центр МГУ

Защита состоится 5 июня в 15 ч. 00 мин. на заседании Диссертационного совета ДМ003.032.01 в Институте систем информатики имени А.П.Ершова Сибирского отделения РАН по адресу:
630090, г. Новосибирск, пр. Акад. Лаврентьева, 6.

С диссертацией можно ознакомиться в читальном зале ИСИ СО РАН
(г. Новосибирск, пр. акад. Лаврентьева, 6).

Автореферат разослан «4» мая 2009 г.

Ученый секретарь
Диссертационного совета,
к.ф.–м.н. Мурзин Ф.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В связи со стремительным ростом объемов текстовой информации в электронных хранилищах данных, включая Интернет, возникает потребность в разработке *человеко-машинных интерфейсов*, а также *систем автоматического извлечения фактов и знаний* из текстов различной языковой природы. Серьезные проблемы при этом создает *вариативность* языковых единиц (ЯЕ), проявляющаяся в разных формах на всех уровнях иерархии. Для *автоматического обнаружения и отождествления* вариантов ЯЕ в тексте необходимо разрабатывать специальные программные средства с опорой на алгоритмы, *формализующие понятие* ЯЕ *и ее ближайшей окрестности*, что и определяет направленность данной работы. В основе таких алгоритмов лежит изучение закономерностей варьирования ЯЕ, в частности, выявление допустимых преобразований (редакционных операций), а также получение *количественных характеристик вариативности* ЯЕ. Они включают в себя формальные оценки близости двух ЯЕ, доминирующие типы редакционных операций, оценки устойчивости различных позиций внутри ЯЕ, характеризующие возможности ее членения на более мелкие единицы, и др. Сложность автоматизации исследования вариативности состоит в том, что программный комплекс должен включать широкий набор средств автоматической обработки текстов (АОТ), в частности, программы фильтрации вариантов, не представляющих интереса (словоизменение) и др.

Традиционные лингвистические исследования вариативности ЯЕ носят преимущественно качественный характер. *Отличительными особенностями* данной работы, проводимой на материале русского языка и отдельных его подъязыков, являются *количественный характер исследования* и *ориентация не только на единицы низких уровней* (корни и канонические формы слов), *но и более высоких − слабо формализованных* (устойчивые словосочетания, фразы, построенные на «игре слов», индикаторы отдельных аспектов содержания, сверхфразовые единства и т.п.). Эти особенности определяют широкую сферу

применимости программно-алгоритмического аппарата, созданного для анализа вариативности ЯЕ. Он может быть использован для *обнаружения дубликатов*, заимствований (в том числе в текстах программ), *оптимизации информационного поиска* (обогащение запроса путем варьирования, формирование шаблонов), *сегментации* длинных неструктурированных текстов, *обнаружения ошибок* и *стилеметрии* (формализация понятий «норма», «отклонение от нормы»).

Цели исследования: 1) разработка и программное обеспечение методики выделения и количественного анализа вариативности ЯЕ; 2) исследование закономерностей варьирования ЯЕ на разных уровнях иерархии и использование этих закономерностей в приложениях.

Основные направления исследований:

− разработка алгоритмов и программ предобработки текстов и выделения ЯЕ более высокого уровня, чем слово;

− количественное изучение вариативности на уровне морфем и лексем русского языка; формирование и использование электронного словаря паронимов;

− анализ вариативности словосочетаний и фраз (на материале газетных заголовков, построенных на «игре слов», и аспектных маркеров); использование полученных результатов при создании программ обогащения индикаторных (аспектных) словарей и построения квазирефератов текста;

− анализ вариативности взаимосвязанных текстов (дублирующие переводы; квазирефераты одного и того же текста, построенные разными программами).

Методы исследований опираются на межъязыковые аналогии, аппарат *L*-граммного представления текстов, используют технику динамического программирования для сравнения символьных объектов, элементы математической статистики и теории формальных языков, принципы структурного и модульного программирования.

На защиту выносится:

− методика количественного исследования проявлений вариативности ЯЕ разных уровней иерархии, реализованная в виде совокупности методов и программ

выделения ЯЕ из текста, их нормализации, построения ближайших окрестностей ЯЕ и их количественной характеризации;

− результаты апробации методики на разных типах ЯЕ с иллюстрацией возможностей использования в реальных системах обработки текстов (информационный поиск, обнаружение ошибок, построение квазирефератов и др.).

Научная новизна:

− *впервые* получены *оценки комбинаторной вариативности корней, слов, морфемных моделей*, численно характеризующие процессы словообразования в русском языке. Результаты представлены в виде *электронного словаря паронимов* (графемный и фонемный варианты), зарегистрированного во Всероссийском научно-техническом информационном центре (ВНТИЦ, № 50200801785);

− создана *уникальная база данных*, содержащая газетные заголовки, построенные на «игре слов», их прототипы (крылатые фразы, цитаты, пословицы и пр.), а также информацию об авторах и изданиях. *Количественная характеризация схем варьирования прототипов* (в том числе *не упоминавшихся ранее*) дает возможность *устранения штампов*, *проведения цитатно-стилистической экспертизы*, *поиска* подходящего *прототипа* для заголовка.

− предложен и реализован *новый* (*более чувствительный*) *алгоритм выделения сверхфразовых единств* в тексте, основанный на использовании *сканирующих статистик*. Введено понятие *профиля кластеризуемости текста*, аккумулирующее информацию обо всех выделенных в тексте сверхфразовых единствах и позволяющее строить *различные варианты квазирефератов текста* на основе совместного учета частотной и позиционной информации;

− *впервые* проиллюстрирована возможность использования количественных характеристик *L*-граммного спектра для *частичной автоматизации процедуры формирования и обогащения* (путем варьирования) *индикаторных словарей*, фиксирующих подсказки о различных аспектах содержания научного текста.

Достоверность и обоснованность предлагаемых решений подтверждается хорошей корреляцией экспертных оценок с результатами, получаемыми с

помощью разработанных и программно реализованных методов.

Практическая ценность проведенных исследований состоит в том, что созданный *комплекс программ*, *реализующих методику количественного исследования вариативности* ЯЕ, позволяет сформировать описание ЯЕ как совокупности ее возможных вариантов, включая и не представленные в обучающей подборке. Это повышает эффективность информационного поиска и обработки естественноязковых текстов в автоматическом режиме. На основе разработанной методики *построены* многоцелевой *электронный словарь паронимов* русского языка, трудные *тестовые словари* для *систем распознавания* и *синтеза речи*, *индикаторные словари* для многоаспектного анализа научных текстов, являющиеся компонентами баз знаний систем АОТ.

Апробация работы. Основные результаты докладывались на Сибирском конгрессе по прикладной и индустриальной математике (ИНПРИМ-2000); Международных научно-практических конференциях (KDS-2001, 2005); Всероссийской научной конференции "Квантитативная лингвистика: исследования и модели" (КЛИМ-2005); пяти конференциях "Компьютерная лингвистика и интеллектуальные технологии" (Диалог-2003 − Диалог-2007). Многие работы прошли экспертизу в ходе выполнения проектов, поддержанных грантами РФФИ (№ 00-06-80420, 03-06-80118, 06-06-80467) и РГНФ (№ 99-04-12026-в).

Личный вклад. Методика количественного анализа вариативности ЯЕ разработана совместно с руководителем. Основные результаты по исследованию вариативности ЯЕ разного уровня, созданию тестовых и индикаторных словарей, формированию квазирефератов получены автором лично. Выделение сверхфразовых единств с помощью сканирующих статистик реализовано при участии Мирошниченко Л.А.

Публикации. По теме диссертации опубликовано 28 работ: 4 статьи в рецензируемых журналах, 13 − в научных сборниках, 11 − в трудах международных и всероссийских конференций.

Структура работы. Диссертационная работа состоит из введения, обзор-

ной главы 1, четырех глав с изложением основных результатов, заключения и списка литературы (143 наименования), содержит 4 рисунка и 17 таблиц. Общий объем работы − 184 стр.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе представлен обзор работ, связанных с лингвистической трактовкой понятия вариативности и ее проявлениями в разных языковых системах. Рассмотрены формальные меры сходства символьных объектов, с помощью которых можно оценивать близость ЯЕ.

В работах лингвистов дается толкование вариативности в узком и широком смысле. Первое предполагает, что различные по форме варианты ЯЕ сохраняют смысловую близость. Зачастую такое толкование является ограничительным. Так, при обнаружении ошибок паронимического типа (*тест − текст*, *частый − частный*) вариантами слова удобно считать формально близкие, но в общем случае отличающиеся по смыслу ЯЕ. Порой смысловое тождество намеренно нарушается во фразах, построенных на «игре слов»: с*вято кресло пусто не бывает*; *пусто место свято не бывает*; *свято место теперь пусто*.

Широкая трактовка вариативности предложена М.М. Маковским в его монографии «Лингвистическая комбинаторика». Предполагается, что изменение формы ЯЕ может привести к изменению смысла, а трансформация смысла − повлечь за собой преобразование формы ЯЕ. В этом понимании варьирование − одно из средств развития и пополнения языка новыми ЯЕ.

Лингвистические исследования вариативности языка объясняют и систематизируют это явление лишь на качественном уровне. Для получения количественных характеристик вариативности нужно уметь оценивать сходство между символьными объектами на разных уровнях иерархии. Для наших целей известные меры сходства удобно разделить на учитывающие порядок следования элементов в сравниваемых последовательностях и игнорирующие его.

В мерах первого типа фиксируется множество допустимых редакционных операций, отражающих возможности трансформации объектов. Универсальны-

ми элементарными операциями являются замена, вставка и устранение символа. Метрика Левенштейна определяется как минимальное число операций указанного типа, переводящих одну последовательность в другую. Модификации этой метрики (редакционное расстояние и др.) связаны с изменением или расширением состава операций и введением весов для них.

Теоретико-множественные меры сходства не учитывают порядок следования элементов в тексте. Они работают с набором признаков, вычисляемых для каждого текста (это может быть множество *L*-грамм, см. главу 2). Подобные меры вычисляются проще, чем редакционное расстояние. На разных уровнях языковой иерархии используются разные меры сходства.

Анализ проявлений вариативности в других языковых системах (биомолекулы, язык песен, цепные письма и др.) существенно расширяет спектр специфических редакционных операций. Отмечена важность межъязыковых аналогий и возможность переноса отдельных постановок и методов решения из одной языковой системы в другую.

Во второй главе описана методика количественного исследования вариативности ЯЕ, включающая: 1) *формирование обучающих подборок*, содержащих образцы изучаемых ЯЕ и их варианты; 2) *процедуры предобработки текста* (фонетическая транскрипция, морфологический, *L*-граммный и позиционный анализ); 3) *методы выделения из текста* ЯЕ более высокого уровня, чем слово; 4) *анализ допустимых редакционных операций* и выбор мер близости; 5) *определение ближайших окрестностей* каждой ЯЕ и их *количественная характеризация*. Для разных классов ЯЕ некоторые этапы могут носить факультативный характер. Кратко охарактеризуем пункты 1 − 5 методики.

1) *Создание обучающих подборок*. Для анализа корней и слов русского языка выбран деривационный словарь Д. Уорта объемом порядка 100 тыс. слов (свыше 10 тыс. корней), в котором слова имеют межморфемные разделители, что удобно для выделения корней и построения морфемных моделей. Изучение вариативности ЯЕ более высоких уровней проводится на отдельных подмноже-

ствах языка: словосочетаниях, несущих информацию о различных аспектах содержания научных текстов; крылатых выражениях и фразах, используемых в качестве прототипов газетных заголовков (подборка из 2,5 тыс. заголовков). Материалом для изучения вариативности на высших уровнях являлись переводы одного и того же текста или его квазирефераты, сделанные разными людьми или разными компьютерными программами.

2) *Процедуры предобработки текста* применяются к ЯЕ типа «слово» и выше. *Транскрибирование* (представление слов в алфавите фонем) ориентировано на использование в речевых человеко-машинных интерфейсах. Реализованная автором процедура транскрибирования работает с фразами, т.е. со слитно произносимыми словами, что потребовало детального учета взаимовлияния звуков на стыках слов.

С помощью *морфологического анализа* словоформы представляются в каноническом виде. Это необходимо, когда морфологическая вариативность ЯЕ является мешающим фактором (например, при выделении устойчивых словосочетаний). Особенностью реализованной в работе процедуры морфологического анализа, отличающей ее от известных аналогов, является выбор базового словаря, содержащего информацию о морфемной структуре слова. Это позволяет анализировать вариативность сразу на двух уровнях − морфемном и лексемном, прослеживать взаимосвязи между ними, проводить межъязыковые аналогии (см. главу 3). Алгоритм предусматривает нормализацию «новых» (не содержащихся в базовом словаре) слов, используя рассуждения «по аналогии» и информацию о разнообразии форм «нового» слова в исследуемом тексте.

L-граммный анализ − это способ представления текста в виде набора цепочек из L подряд следующих букв (на нижних уровнях) или слов (на верхних) с указанием частоты встречаемости и мест вхождения их в текст. Совокупность всевозможных содержащихся в тексте L-грамм с сопутствующей информацией образует частотную характеристику текста порядка L, обозначаемую $\Phi_L(T)$. Совокупность $\Phi_L(T)$ со значениями L от 1 до L_{max} (длина в символах или словах

максимального повтора в тексте) составляет полный частотный спектр текста Ф(T). Он используется для выявления устойчивых словосочетаний, максимальных повторов (структурных единиц достаточно высокого уровня), обнаружения ошибок, изучения особенностей авторского стиля.

Аналогом Φ_L (T) для группы текстов $\overline{T} = (T_1, T_2, \ldots, T_m)$ является совместная частотная характеристика L-го порядка $\Phi_L(\overline{T})$, содержащая частотную и позиционную информацию об L-граммах, общих хотя бы для пары текстов из $\overline{T}$. Совокупность $\Phi_L(\overline{T})$ со значениями L от 1 до $L_{max}(\overline{T})$ (длина максимального *межтекстового* повтора) образует совместный частотный спектр группы текстов − $\Phi_L(\overline{T})$. Он используется для выделения отдельных классов ЯЕ (в частности, аспектных маркеров) и лежит в основе вычисления теоретико-множественных мер близости для пар и групп текстов. Для вычисления полных частотных спектров используются «trie-структуры» (L-граммные деревья). Трудоемкость алгоритмов имеет порядок $L_{max} \cdot N$, где L_{max} − длина максимального внутри- (или меж-) текстового повтора, N − длина текста (или группы текстов).

Позиционный анализ оценивает значимость ЯЕ на основе информации о местах вхождения ее в текст. Предполагается, что наиболее значимыми являются ЯЕ, распределенные по тексту неравномерно, в частности, кластеризованные ЯЕ. Для их обнаружения адаптирован аппарат сканирующих статистик, характеризующийся наибольшей чувствительностью к такого рода аномалиям. С его помощью удается выявлять и ряд других аномалий, в частности, сверхравномерно распределенные по тексту ЯЕ (потенциальные разделители).

3) *Методы выделения* ЯЕ более высокого уровня, чем слово, основаны на аппарате L-граммного и позиционного анализа. Рассмотрены три типа ЯЕ: *устойчивые словосочетания*, *индикаторы* отдельных аспектов содержания научного текста (*аспектные маркеры*) и *сверхфразовые единства*. Устойчивые словосочетания доминируют в словарях терминологической лексики и служат универсальной базой для выделения других типов структурных единиц. Аспектные маркеры являются перспективным инструментом информационного поиска,

однако формирование словаря этих ЯЕ под новый аспект содержания, как правило, производится вручную. Даже частичная автоматизация этого процесса представляется актуальной. Сверхфразовые единства − пример более крупных ЯЕ, определяющих макроструктуру текста, что существенно при работе с неструктурированными документами, характерными для сети Интернет.

Термином «*устойчивая цепочка*» мы характеризуем L-граммы ($L \geq 2$), встречающиеся в большом числе различных контекстов. Максимально неустойчивой считается цепочка, которая лишь единственным образом продолжаема в обе стороны. Это означает, что она не имеет самостоятельного значения и функционирует лишь в составе более длинной цепочки. Формально, пусть x − произвольная L-грамма, $F(x)$ − частота ее встречаемости в тексте. Из всех левосторонних расширений x, реализованных в тексте и имеющих форму ax (a − произвольная словоформа, предшествующая x), выберем расширение a^*x с максимальной частотой встречаемости в тексте. Очевидно, что $F(a^*x) \leq F(x)$. Аналогично, среди всех правосторонних расширений вида xb выберем самое частое − xb^*, при этом $F(xb^*) \leq F(x)$. Цепочка с $F(x) \geq 2$ считается *устойчивой*, если одновременно выполняются условия: $F(a^*x)/F(x) \leq П$ и $F(xb^*)/F(x) \leq П$, где значение порога $П$ не превышает 0,5. Такой выбор порога устраняет возможность доминирования по частоте любого из возможных расширений.

Для выявления *аспектных маркеров* используется гипотеза об их устойчивой повторяемости в разных текстах. В отдельно же взятом тексте конкретный маркер не должен встречаться более одного- двух раз, поскольку основные аспекты содержания (цель, актуальность, новизна, ...) обычно формулируются однократно. Исходя из этого, потенциально возможные аспектные маркеры мы ищем среди нормализованных устойчивых L-грамм из $\Phi_L(\overline{T})$, удовлетворяющих условию $F_{абс}(x)/F_{текст}(x) \leq 2$, где $F_{абс}(x)$ − число вхождений L-граммы x в тексты из $\overline{T}$, а $F_{текст}(x)$ − число текстов из $\overline{T}$, содержащих x. Эксперт осуществляет дополнительную фильтрацию отобранных L-грамм с привлечением минимального контекста (1−2 предложения). Прочтение всех текстов из $\overline{T}$ для отбо-

ра маркеров вручную требует гораздо большего времени.

Сверхфразовые единства − это достаточно крупные фрагменты, связующими элементами в которых выступают кластеризованные знаменательные словоформы. Для выявления кластеров используется статистика $d(n)$, равная длине минимального фрагмента, содержащего ровно n вхождений нормализованной словоформы x ($n_{пор} \le n \le F(x)$, где $F(x)$ − частота встречаемости словоформы в тексте, а $n_{пор}$ − ограничение снизу на число повторов в кластере). Распределение $d(n)$ при нулевой гипотезе в непрерывном случае (точки на отрезке) известно и частично затабулировано. В нашем случае аномалии в распределении лексем в тексте фиксируются с помощью имитационного моделирования. Кластеризация имеет место, если: $(S_{набл} \le S_{min})\&\ (S_{набл} \le \overline{S} - 3s)$, где $S_{набл}$ − наблюдаемое значение статистики $d(n)$ в анализируемом тексте, а S_{min} и $\overline{S}$ − соответственно, минимальное и среднее значения статистики $d(n)$, полученные в серии из 100 экспериментов с «рандомизированными» текстами, s − среднеквадратичное отклонение. Рандомизация проводилась путем равномерного перемешивания словоформ исходного текста.

4) *Анализ допустимых редакционных операций* и выбор мер близости осуществляются на основе обучающих подборок, содержащих примеры ЯЕ и их вариантов. Показательна в этом плане подборка газетных заголовков, построенных на «игре слов», где одному прототипу (инварианту) может соответствовать до 10÷15 вариантов, а число зафиксированных схем варьирования близко к 30. При исследовании вариативности корней, слов, морфемных моделей слов используются операции вставки, замены, устранения символа и метрика Левенштейна для сравнения ЯЕ. При исследовании средних уровней используем операции вставки, замены и устранения слов. Сравнение ЯЕ высокого уровня проводим на основе теоретико-множественных мер сходства, адекватно реагирующих на дупликации и перестановки крупных блоков в тексте.

5) *Определение ближайших окрестностей* ЯЕ и их количественная характеризация − наиболее трудоемкий этап методики. Поясним его на примере

получения ближайших окрестностей слов. Пусть V − исходный словарь ЯЕ, $d(a, b)$ − редакционное расстояние между a и b ($a, b \in V$). Если веса операций одинаковы и равны 1, d может принимать значения 0, 1, 2, … Совокупность всех ЯЕ из V, удаленных от a ($a \in V$) не более чем на d, назовем d-окрестностью a и будем обозначать $v_d(a)$. Например, для $a = порт$ и $d = 1$ $v_1(a) = \{апорт, спорт, пот, орт, борт, корт, сорт, торт, форт, хорт, пора, пост, поэт\}$. Если d мало, пары (a, b), где $b \in v_d(a)$, трактуются как паронимы в широком смысле.

Если ввести обозначения S, I и D для операций замены (Substitution), вставки (Insertion) и устранения символа (Deletion), то полную окрестность $v_1(a)$ можно представить в виде $v_1(a) = v^S(a) \bigcup v^I(a) \bigcup v^D(a)$, где $v^S(a)$, $v^I(a)$, $v^D(a)$ − наборы ЯЕ из V, отличающихся от a, соответственно, одной заменой, вставкой или делецией. В свою очередь, $v^S(a) = \bigcup_k v_k^S(a)$, где $a = a_1a_2\ldots a_k\ldots a_j$, $1 \le k \le j$, $v_k^S(a)$ − совокупность слов, отличающихся от a только заменой по k-ой позиции. Аналогично, $v^D(a) = \bigcup_k v_k^D(a)$, где $v_k^D(a)$ − либо одноэлементные ($v_k^D = \{a_k\}$), либо пустые множества. В случае вставок $v^I(a) = \bigcup_k v_k^I(a)$, при этом индекс k меняется от 1 (вставка перед a_1) до $j + 1$ (вставка после a_j) . Символы, замещающие k-ю позицию в ЯЕ, составляют *векторы замен* $sub_k(a)$, *вставок* $ins_k(a)$ и *делеций* $del_k(a)$. Чем меньше длина вектора, тем устойчивее данная позиция.

Задача построения вариантов ЯЕ эквивалентна отысканию несовершенных повторов. Для $d = 1$ она сводится к более простой задаче отыскания точных повторов путем использования специальных "склеивающих" преобразований, делающих неразличимыми слова, отличающиеся друг от друга только заменой или вставкой/делецией по k-й позиции.

При $d = 2$ возможны следующие комбинации искажений: *SS*, *II*, *DD*, *SI*, *SD*, *ID*. Аналогично случаю $d = 1$ для любого $a \in V$ определяются подокрестности $v^{\alpha\beta}(a)$ для каждой из комбинаций $\alpha, \beta \in \{S, D, I\}$, а также соответствующие

им векторы искажений. Поиск соседей облегчается, если словарь V разделен на подмножества V_j слов длины j ($V = \bigcup V_j$, j = 1,2,…). В зависимости от комбинации α, β сравниваются только элементы множеств V_j (схемы *SS* и *ID*); V_j и V_{j+1} (схема *SI*); V_j и V_{j-1} (схема *SD*); V_j и V_{j+2} (схема *II*); V_j и V_{j-2} (схема *DD*).

Для ЯЕ высокого уровня ближайшие окрестности могут быть сформированы лишь частично (по ограниченному подъязыку, ограниченному набору операций, ограниченной обучающей подборке). Но даже в этом случае удается выявить допустимые редакционные операции и сформировать поисковые шаблоны для ЯЕ, учитывающие возможные проявления вариативности.

Кроме описанных выше алгоритмов предобработки, выделения ЯЕ и построения окрестностей реализован также ряд процедур, иллюстрирующих возможности практического использования разработанного аппарата. Это процедуры формирования тестовых словарей для систем распознавания и синтеза речи (см. гл. 3), а также построения профиля кластеризуемости и квазирефератов текста (см. гл. 5). На рис. 1 показаны схемы сборки программных модулей обработки текста для получения конкретного продукта (словаря, квазиреферата, графика, позволяющего осуществить сегментацию текста). Для иллюстрации на рисунке двойными стрелками изображен процесс получения тестовых речевых словарей, а жирными − построения квазирефератов.

В третьей главе исследуется вариативность корней ($\overline{a}$) и слов (a). Последние рассматриваются на фонемном, графемном и морфемном уровне. Указаны возможности практического использования полученных результатов.

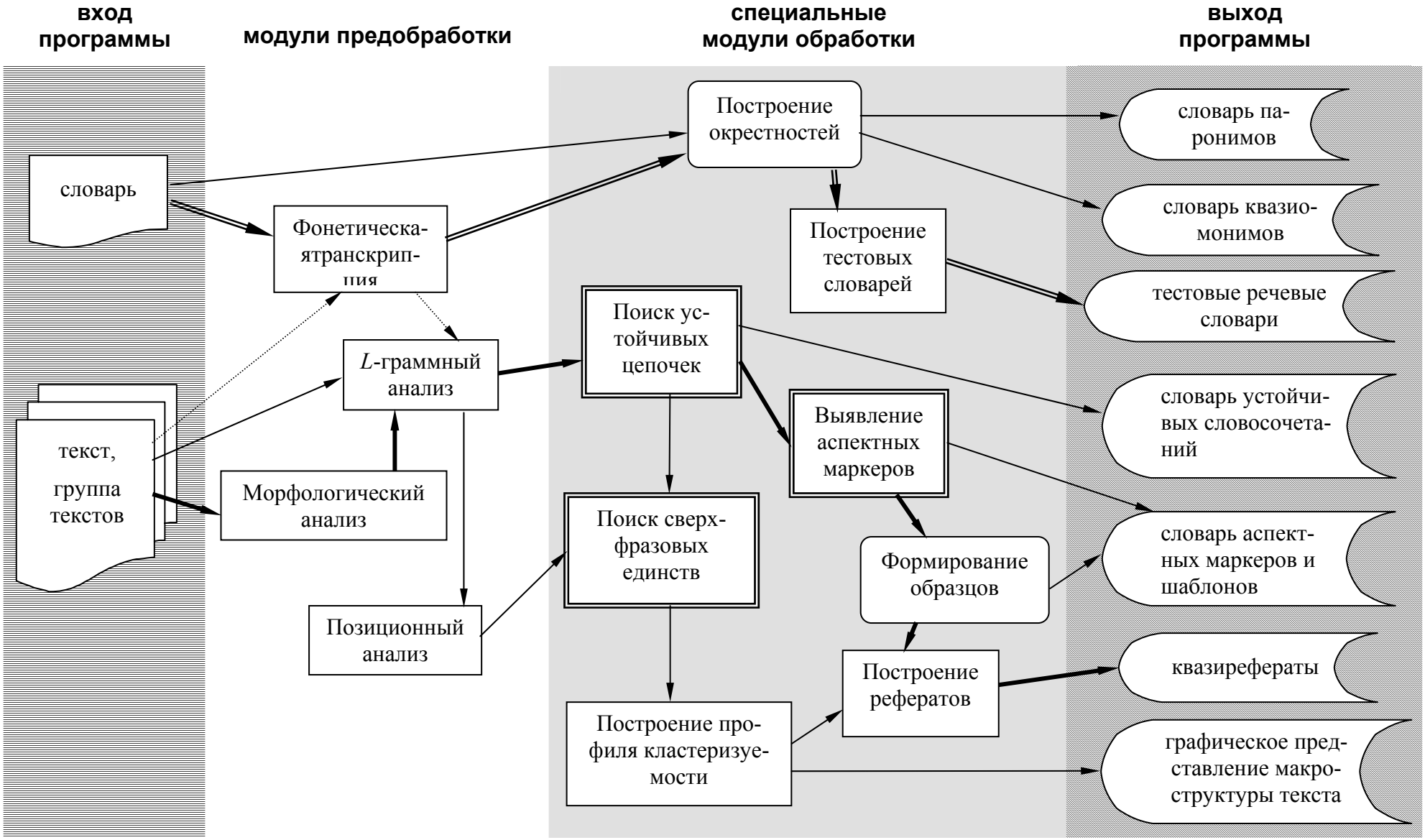

Рис. 1. Схема сборки модулей обработки текстовых данных

Непустые 1-окрестности имеют при использовании всех трех операций (*S*, *I*, *D*) примерно 43% слов и 61% корней, т.е. степень проявления паронимии достаточно высока. Превалирующий тип искажений − *замены*. Наиболее вариативны *короткие* слова и корни. Приведем примеры корней- и слов-рекордистов: *а* = *полить*, $sub_1(a)$={*д*, *з*, *м*, *п*, *с*, *х*}, $sub_2(a)$ ={*а*, *и*, *о*, *ы*, *я*}, $sub_3(a)$ ={*б*, *в*, *ж*, *л*, *п*, *ч*, *ш*}, $sub_4(a)$ ={*и*, *о*, *с*}, всего 17 соседей; $\bar{a}$ = *мал*, $sub_1(\bar{a})$={*б*, *в*, *г*, *д*, *ж*, *з*, *к*, *л*, *м*, *п*, *с*, *т*, *ф*, *х*, *ч*, *ш*}, $sub_2(\bar{a})$ ={*а*, *г*, *е*, *и*, *о*, *у*, *ю*}, $sub_3(\bar{a})$={*в*, *г*, *д*, *ж*, *з*, *й*, *к*, *л*, *м*, *н*, *р*, *с*, *т*, *ф*, *х*, *ц*, *ч*, *ш*, *щ*}, всего 39 соседей.

Вариативность разных позиций в слове/корне существенно отличается. *Существуют доминирующие типы вставок и замен* для разных (но не всех) позиций слов. Обычно они приходятся на *начальные* (k = 1, 2, 3) и, в меньшей степени, *конечные* ($k = j$-2, j-1, j, j+1) *позиции*. Случаи доминирования тесно связаны с морфемной структурой и проявляют себя при значительных длинах слов ($j \geq 6$). В корнях явное доминирование одних типов искажений над другими чаще всего объясняется чередованием гласных и согласных. Векторы замен и вставок обычно однородны по СГ-составу (С − согласный, Г − гласный).

При переходе к фонетической записи ближайшие окрестности слов могут измениться. Так, в графемном варианте *d*(*явить*, *свить*) = 1, а в фонетическом − *d*(J^в’ит’, с’в’ит’) = 2, тогда как *d*(*явить*, *ловить*) = 2, но *d*(J^в’ит’, л^в’ит’) = 1. Для построения тестовых речевых словарей выбираются слова, у которых несовпадающие фонемы близки по артикуляционно-акустическим характеристикам, например, [т’, к’]: *тенор* − *кенар*; [м, н]: *исламский* − *исландский* и т.п. Реализован алгоритм формирования тестовых словарей приемлемых объемов (от 50 до 100 пар слов) с возможностью многократного обновления, что затрудняет настройку тестируемой речевой системы на конкретный словарь.

При d = 2 уже порядка 82% всех слов имеют непустую 2-окрестность. Комбинации допустимых операций ранжируются по частоте встречаемости следующим образом: *SS* > *SD* > *SI* > *DD* > *II* > *ID*. Слова с пустой 2-окрестностью можно отнести к категории *устойчивых к искажениям* парони-

мического типа (*взгляд*, *соблазн*, *ремесло*). Схемы *SS* и *ID* при сохранении буквенного состава трактуются либо как *перестановки* символов (*SS*: *теплица* − *петлица*), либо как *транспозиции* (перенос символов): *укорять* − *рукоять* (*ID*).

Важное значение приобретает распределение искажений по позициям слова. Будем различать *кластеризованные* искажения (они затрагивают соседние позиции) и *некластеризованные* (разнесенные друг от друга). Показано, что *число первых значительно выше* уровня, допускаемого моделью с независимым распределением искажений в слове. *Искажение соседних позиций* служит *индикатором структурной единицы* более *низкого уровня* (слога, морфемы). *Связь эффекта кластеризации* со структурой ЯЕ может быть постулирована для других языковых систем и использована для выделения ЯЕ из слитных текстов (например, генетических).

	M	*m'*	*n*	*r*
1	2471	*pRssf*	8947	3
2	1680	*pRsf*	14105	1
3	1365	*pRsssf*	2155	11
4	1058	*Rssf*	7181	4
5	955	*Rsssf*	3404	8
6	896	*ppRsf*	1380	22
7	790	*ppRssf*	827	15
8	693	*pRssssf*	1010	19
9	522	*pRf*	1392	14
10	461	*pRs*	1447	13

Вариативность морфемных моделей рассматривалась для $d = 1$. Простейшие модели описывают группы слов в виде цепочек морфем, в которых корневая морфема унифицирована. Например, модель *m* = под-*R*-к-а описывает слова с $R \in \{$*бор*, *вод*, *зем*,...$\}$. Более высокий уровень агрегирования имеют *типовые модели*, где кроме корня унифицированы еще все префиксальные (*p*) и суффиксальные морфемы (*s*, *f*, *c*). Так, типовая модель *m'* = *pRsfc* описывает множество простейших моделей: m_1 = под-*R*-и-ть-ся, m_2 = рас-*R*-а-ть-ся, m_3 = вы-*R*-я-ть-ся.

В приводимой таблице указаны первые десять типовых моделей *m'*, упорядоченных согласно числу *M* охватываемых ими простейших моделей (*m*). Здесь *n* − число слов словаря, описываемых типовой моделью *m'*, *r* − ранг типовой модели при упорядочении по *n*. Получены оценки вариативности морфемных моделей, позволяющие количественно охарактеризовать процессы словообразования в русском языке, а также выявить «незаполненные» позиции (отсутствую-

щие в языке формы слов). Результаты данной главы могут быть использованы также для обнаружения *ошибок* паронимического типа, составления *лингвистических задач*, тестирования систем распознавания и синтеза речи.

В четвертой главе приведены результаты количественного анализа вариативности словосочетаний (аспектные маркеры) и фраз (газетные заголовки). Рассмотрена возможность моделирования вариантов по прототипам.

Аспектные маркеры (слова, словосочетания, шаблоны) применяются для автоматического извлечения информации о различных аспектах содержания текста. Например, аспект «цель исследования» выявляется с помощью маркеров типа «в настоящей работе», «в работе рассматривается», «целью является» и др. По обучающей подборке трудов конференции Диалог'2002 с помощью алгоритма, описанного в главе 2, построены индикаторные словари для выявления 12 аспектов содержания (цель работы, актуальность, новизна, полученные результаты и др.). Суммарный объем словарей по всем 12 аспектам составил порядка 700 маркеров. Анализ маркерных цепочек позволяет выделить группы условно синонимичных подстановок. Так, наличие маркеров «в данной работе» и «в настоящей работе» позволяет считать слова «данный» и «настоящий» условными синонимами в контексте рассматриваемого аспекта. Аналогичный вывод можно сделать относительно глаголов «рассматриваться» и «обсуждаться» («в работе рассматривается», «в работе обсуждается»…). Формирование групп условных синонимов типа X = {статья, доклад, работа,…}, Y = {данный, предлагаемый, настоящий,…}, Z = {рассматриваться, обсуждаться, описываться,…} дает возможность обогащения исходного словаря путем варьирования уже отобранных маркеров, т.е. без пополнения обучающей подборки. При этом исходные маркеры заменяются шаблонами вида: цель\$x$; в\$y$\статье; в\работе\$z$ и т.п., где переменные x, y, z допускают подстановки из элементов множеств X, Y, Z соответственно. Используя эти и другие типы варьирования, суммарный объем индикаторных словарей был доведен примерно до 1000 маркеров. Эксперименты на контрольной подборке показали приемлемую полноту (~ 80−90%) и точ-

ность (~ 65−85%) идентификации аспектов. Реализован алгоритм построения квазиреферата научной статьи по заданному набору аспектов.

На материале газетных заголовков, построенных путем варьирования общеизвестных прототипов, получены *качественные и (впервые) количественные (частотные) характеристики прототипов*, *их источников* и *схем варьирования*. Замена слова в прототипе − наиболее частая операция (28,8% всех случаев): «Пролетая над гнездом науки». Интересными являются схемы варьирования, обыгрывающие *многозначность* ЯЕ (1,6%): «Шаром покати» − статья о боулинге; *фонетическое сходство* (6,7%): «Все течет, все измеряется»; использующие *префиксные и суффиксные блочные делеции* (6,3%): «Служить бы рад»…, «…Табачок врозь»; *контаминации* (1%): «Красному петуху море по колено» − о пожаре в сауне. Часто используются *согласованные двойные замены* (5,7%) с сохранением синтаксической структуры прототипа: «Место преступления определить нельзя». Комбинация различных преобразований имеет место в 18,8% случаев («В спорах о гимне рождается мелодия»). Многократное использование одного типа преобразования, например, антонимического («Новый враг хуже старых двух»), встречается редко.

Результаты этой главы представляют интерес в плане *изучения специфики варьирования* ЯЕ на разных уровнях иерархии, устранения штампов, повышения *эффективности информационного поиска* (варьирование запроса), выработки подходов к *автоматизации отдельных схем варьирования*.

В пятой главе анализируются структурные единицы верхнего уровня (варьированные тексты), полученные путем перевода одного и того же текста на другой язык или его реферирования разными людьми или программами.

Сравнивались два перевода на русский язык книги Алана А. Милна «Винни-Пух». Ранний сделан Б. Заходером (З), более поздний − В. Вебером (В) и Н. Рейн. Оценки сходства и различия этих текстов получены на основе анализа совместного частотного спектра. Показана особая роль «контрастных» (т.е. представленных преимущественно в одном из текстов) *L*-грамм в выявлении

композиционных и стилистических различий двух переводов, а также проявлений целенаправленного варьирования оригинала. К ним можно отнести *переименование действующих лиц* (к этому прибегают и Заходер, и Вебер), *русификацию системы мер и весов* (З), *вольный перевод звукоподражаний*, *восклицаний* (З и В), замену часто встречающегося слова группой условных синонимов (В) и наоборот.

Отмечен весьма специфический вид варьирования, сводящийся к сознательному дистанцированию от имеющегося известного перевода. А именно, в тех местах, где Заходер почти дословно следует Милну, Вебер варьирует его. Там же, где Заходер отходит от оригинала, Вебер следует Милну или, в свою очередь, отходит от оригинала. Но Заходер сам предупреждает о возможности отклонений от оригинала («пересказал Борис Заходер»), тогда как Вебер настаивает на близости к оригиналу («ничего не привносить своего»).

В качестве вариантов текста можно рассматривать различные его свертки в виде квазирефератов. Предложены два способа формирования квазиреферата на основе позиционно кластеризованных лексем. Первый связан с построением *профиля кластеризуемости лексических единиц в тексте*. Он отражает совокупное распределение в тексте и взаимосвязь кластеризованных ЯЕ (слов и словосочетаний). Формально *профиль кластеризуемости* − это ступенчатая функция, аргументом которой является порядковый номер предложения в тексте, а значение равно числу различных кластеров, включающих в себя данное предложение. На приводимом ниже рисунке изображен фрагмент профиля кластеризуемости главы 6 из «Винни-Пуха». Ось абсцисс направлена вниз, а ось ординат − по горизонтали слева направо. Вместо значений функции для наглядности выписаны ЯЕ, кластеризованные в данном фрагменте.

номера фраз	*число кластеров*	*кластеризованные* ЯЕ
1 ÷ 54	0	—
55 ÷ 101	2	СЕГОДНЯ; ДЕНЬ РОЖДЕНИЯ;
102 ÷ 110	0	—
111 ÷ 148	1	ГОРШОЧЕК;
149 ÷ 150	5	ГОРШОЧЕК; ПОПРОСИТЬ; НАПИСАТЬ; ДЕРЖАТЬ; ХОТЕТЬ;

151 ÷ 167	6	ГОРШОЧЕК; ПОПРОСИТЬ; НАПИСАТЬ; ДЕРЖАТЬ; ХОТЕТЬ; СОВА;
168 ÷ 169	5	ГОРШОЧЕК; ПОПРОСИТЬ; НАПИСАТЬ; ХОТЕТЬ; СОВА;
170 ÷ 171	4	ГОРШОЧЕК; ПОПРОСИТЬ; НАПИСАТЬ; СОВА;
172 ÷ 178	3	ГОРШОЧЕК; НАПИСАТЬ; СОВА;
179 ÷ 195	2	НАПИСАТЬ; СОВА;
196 ÷ 204	1	СОВА;

Отбор фраз для квазиреферата производится по точкам изменения значений профиля. Это перекликается с позиционным методом реферирования, учитывающим наиболее информативные (начальные и конечные) фрагменты в структуре текста, задаваемой автором. Предлагаемый же метод отталкивается не от явленной структуры (она может быть слишком бедной), а строит независимую оценку макроструктуры текста в виде профиля кластеризуемости. Другой способ построения квазиреферата состоит в приписывании каждому предложению веса в соответствии с наличием в нем кластеризованных словоформ.

Апробация различных подходов к построению квазирефератов демонстрирует многообразие вариантов получаемых решений, что обусловлено специфическими особенностями разных подходов. Так, в рефераты, полученные путем «взвешивания» фраз, могут не попасть короткие, но информативные подзаголовки, поскольку короткие фразы объективно имеют меньше шансов набрать большой вес. Некоторые коммерческие программы не отделяют общеупотребительную лексику от тематической, что ухудшает качество квазиреферата. Служебные слова редко кластеризуются, но когда это случается, они учитываются в профиле кластеризуемости. В рефераты, основанные на индикаторных словарях, могут не попасть информативные фразы, не содержащие аспектного маркера. Эти примеры приводят к выводу, что для получения качественного квазиреферата желательна комбинация различных подходов, адекватно учитывающих широкий спектр проявлений вариативности единиц данного уровня.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ.

1. Предложены и реализованы *новые* алгоритмы выделения в тексте структурных единиц *более высокого уровня*, *чем слово*: 1) устойчивых словосочетаний, 2) маркеров различных аспектов содержания текста, 3) сверхфразовых единств, соотносимых с отдельными микротемами текста.

2. *Усовершенствованы* и реализованы алгоритмы предобработки текстов: 1) транскрипции с расширенным алфавитом фонем, 2) нормализации текста с учетом новых слов, 3) *L*-граммного анализа текста, группы текстов.

3. Создан программный комплекс реализующий методику количественного исследования проявлений вариативности ЯЕ разных уровней иерархии, включающий модули *предобработки* текста, *выделения* ЯЕ *из текста*, и *формирования ближайшей окрестности* ЯЕ и *получения числовых оценок вариативности*.

4. *Впервые* с помощью разработанных программных средств получены *количественные оценки вариативности* ЯЕ разных уровней: корней, слов, морфемных моделей, аспектных словосочетаний, крылатых фраз, параллельных текстов. Показаны возможности использования полученных результатов для: 1) *обогащения* (путем варьирования) *словарей аспектных маркеров*, что существенно *повышает эффективность поиска* отдельных *аспектов содержания* научных текстов; 2) *построения квазирефератов* неструктурированных (в общем случае) текстов путем их сегментации на отдельные микротемы.

5. На базе 100-тысячного словаря русского языка (*V*) построен *уникальный электронный словарь паронимов* «в широком смысле», где каждое слово представлено своими 1-, 2-окрестностями, содержащими слова из *V*, отличающиеся от заданного, соответственно, одним или двумя искажениями типа «вставка», «замена» или «устранение» символа в любой их комбинации. Словарь предназначен для *изучения процессов словообразования*, *поиска и моделирования ошибок* паронимического типа, *генерации комбинаторных лингвистических задач*, *моделирования заголовков*, построенных на «игре слов».

6. Разработана и реализована *методика автоматизированного создания и обогащения* (путем варьирования) *индикаторных словарей*, предназначенных для выявления отдельных аспектов содержания научных текстов. Она позволяет экспертам отбирать аспектные маркеры без прочтения полных текстов и обеспечивает приемлемые результаты по полноте и точности поиска.

Содержание диссертации отражено в следующих работах:

1. Гусев, В.Д. Электронный словарь паронимов: версия 1 / В.Д. Гусев, Н.В. Саломатина // НТИ, серия 2, Информационные процессы и системы. – М.: ВИНИТИ, 2000. − № 6. − С. 34−41.

2. Гусев, В.Д. Электронный словарь паронимов: версия 2 / В.Д. Гусев, Н.В. Саломатина // НТИ, серия 2, Информационные процессы и системы. − М.: ВИНИТИ, 2001. − № 7. − С. 26−33.

3. Загоруйко, Н.Г. Система OntoGrid для автоматизации процессов построения онтологий предметных областей / Н.Г. Загоруйко, …, Н.В. Саломатина // Автометрия. − Новосибирск, 2005. − Т. 41, № 5. − С. 13−25.

4. Гусев, В.Д. Выявление аномалий в распределении лексических единиц по тексту / В.Д. Гусев, Л.А. Мирошниченко, Н.В. Саломатина // Вестник СПбУ, сер. 9. Вып. 3. − Санкт-Петербург, 2005. − С. 64−69.

5. Кельманов, А.В. Правила и алгоритм преобразования орфографической записи на русском языке в фонетическую транскрипцию / А.В. Кельманов, Н.В. Саломатина и др. // Прикладные системы искусственного интеллекта. Вычислительные системы, вып. 153. − Новосибирск, 1995. − С. 32−92.

6. Саломатина, Н.В. Создание тестовых словарей для систем распознавания речи на основе электронного словаря паронимов / Н.В. Саломатина // Квантитативная лингвистика и семантика. Сборник научных трудов. Вып. 2. − Новосибирск, 2000. − С. 63−72.

7. Саломатина Н.В. Создание и исследование компьютерного словаря паронимов / Н.В. Саломатина // Анализ данных и сигналов. Выч. сист., вып.163. − Новосибирск, 1998. − С. 97−112.

8. Гусев, В.Д. Определение и анализ ближайших окрестностей корней слов русского языка / В.Д. Гусев, Н.В. Саломатина // Обнаружение эмпирических закономерностей. Выч. сист., вып.166. − Новосибирск, 1999. − С. 80−103.

9. Гусев, В.Д. Анализ ошибок, не выявляемых автоматическими корректорами / В.Д. Гусев, Н.В. Саломатина // II-я Межвуз. конф. "Квантитативная лингвистика и семантика" (КВАЛИСЕМ-99), тезисы докладов, Новосибирск, 12−15 октября 1999. − НГПУ, 1999. − С. 8−12.

10. Саломатина, Н.В. Количественные характеристики вариативности морфемных моделей / Н.В. Саломатина // Методы обнаружения эмпирических закономерностей. Выч. сист., вып. 167. − Новосибирск, 2001. − С. 93−114.
11. Гусев, В.Д. Количественные исследования вариативности языковых единиц / В.Д. Гусев, Н.В. Саломатина // Труды международной научно-практической конференции KDS-2001. − Санкт-Петербург, 2001. − Том 1. − С. 186−193.
12. Гусев, В.Д. Анализ *L*-граммных словарей параллельных текстов / В.Д. Гусев, Н.В. Саломатина // Труды междунар. конф. Диалог-2003 "Компьютерная лингвистика и интеллектуальные технологии", Протвино, 11−16 июня 2003. − М.: Наука, 2003. − С. 578−582.
13. Гусев, В.Д. Язык заголовков как модель изучения вариативности цитируемых словосочетаний / В.Д. Гусев, Н.В. Саломатина // Лингвистические этюды. Памяти проф. А.М. Моисеева. − Санкт−Петербург, 2004. − С. 203−222.
14. Саломатина, Н.В. Комбинированный алгоритм морфологического анализа для нормализации неизвестных системе слов / Н.В. Саломатина // Анализ структурных закономерностей. Выч. сист., вып.174.− Новосибирск, 2004. − С.61–75.
15. Гусев, В.Д. Алгоритм выявления устойчивых словосочетаний с учетом их вариативности (морфологической и комбинаторной) / В.Д. Гусев, Н.В. Саломатина // Труды междунар. конф. Диалог-2004, Верхневолжский, 2−7 июня 2004. − М.: Наука, 2004. − С. 530−535.
16. Гусев, В.Д. Тематический анализ и квазиреферирование текста с использованием сканирующих статистик / В.Д. Гусев, Л.А. Мирошниченко, Н.В. Саломатина // Труды междунар. конф. Диалог-2005, Звенигород, 1−7 июня 2005. − М.: Наука, 2005. − С. 121−125.
17. Гусев, В.Д. *L*-граммное представление текстов на естественном языке и его возможности / В.Д. Гусев, Н.В. Саломатина // Всерос. научн. конф. Квантитативная лингвистика: исследования и модели (КЛИМ-2005), Новосибирск, 6−10 июня 2005, материалы. − Новосибирск, 2005. − С. 256−270.
18. Гусев, В.Д. Автоматизация формирования индикаторных словарей и возможности их использования / В.Д. Гусев, Н.В. Саломатина // Труды междунар. конф. Диалог-2006, Бекасово, 31 мая − 4 июня 2006. − М: Наука, 2006. − С. 121−125.

19. Гусев, В.Д. Уточнение и обогащение индикаторных словарей для автоматического извлечения информации из научных текстов / В.Д. Гусев, Н.В. Саломатина // Труды междунар. конф. Диалог-2007, Бекасово, 30 мая − 3 июня, 2007. − Москва, 2007. − С. 486−491.

Саломатина Н.В.

МЕТОДЫ И ПРОГРАММНЫЕ СРЕДСТВА ВЫДЕЛЕНИЯ И ЧИСЛЕННОГО ОЦЕНИВАНИЯ ВАРИАТИВНОСТИ ЯЗЫКОВЫХ ЕДИНИЦ

Автореферат

Подписано в печать 29.04.2009 Объем 1,2 уч.-изд. л.
Формат бумаги 60 × 90 1/16 Тираж 100 экз.
Отпечатано в ЗАО РИЦ «Прайс-курьер»
630090, г. Новосибирск, пр. Ак. Лаврентьева, 6, тел. 334-22-02
Заказ №