На правах рукописи

МИХАЙЛОВ Дмитрий Владимирович

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ОЦЕНКИ СЕМАНТИЧЕСКОЙ ЭКВИВАЛЕНТНОСТИ, МОДЕЛИ РАСПОЗНАВАНИЯ И КОМПРЕССИИ ТЕКСТОВ В ОТКРЫТЫХ СИСТЕМАХ КОНТРОЛЯ ЗНАНИЙ

Специальность 05.13.18 – Математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
доктора физико-математических наук

Великий Новгород – 2011

Работа выполнена в ГОУ ВПО “Новгородский государственный университет имени Ярослава Мудрого” на кафедре информационных технологий и систем.

Научный консультант: доктор технических наук, профессор
Емельянов Геннадий Мартинович

Официальные оппоненты: доктор физико-математических наук, профессор
Колногоров Александр Валерианович

доктор технических наук, профессор
Немирко Анатолий Павлович
доктор физико-математических наук
Чернов Владимир Михайлович

Ведущая организация: Научно-исследовательский институт прикладной математики и кибернетики Нижегородского государственного университета им. Н.И. Лобачевского Министерства образования Российской Федерации

Защита диссертации состоится _____ _____ 2011 года в _____ на заседании диссертационного совета **Д 212.168.04** при Новгородском государственном университете имени Ярослава Мудрого по адресу: 173003, г. Великий Новгород, ул. Б. С–Петербургская, д. 41, ауд. _____

С диссертацией можно ознакомиться в библиотеке Новгородского государственного университета имени Ярослава Мудрого.

Автореферат разослан " " 2011 г.

Ученый секретарь диссертационного Совета
Д 212.168.04, кандидат физико-математических наук,
доцент Токмачев М. С.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Тестовое Задание Открытой Формы (ТЗОФ) в системе контроля знаний предполагает ответ испытуемого в виде одного или нескольких предложений Естественного Языка (ЕЯ). Как правило, разработчик теста формулирует свой вариант правильного ответа на основе собственных знаний по заданной Предметной Области (ПО). Традиционно интерпретация ответа испытуемого здесь заключается в простом поиске среди “правильных” вариантов (система “ТестЭкзаменатор”), либо в ручной обработке ответа. Как следует из работ Аванесова В.С., Красильниковой В.А., Майорова А.Н. [1–3], сама постановка ТЗОФ зачастую сводится к простым заданиям на дополнение с ограничениями на ответы. Для правильного оценивания результатов выполнения ТЗОФ необходимо наличие подсистемы обработки ЕЯ, которая, во-первых, способна обрабатывать высказывания с отклонениями от грамматической нормы, а во-вторых, характеризуется единообразием механизмов оперирования предметными и языковыми знаниями. Распространённые модели представления знаний (семантические сети, фреймы и т. п.) характеризуются многообразием способов представления информации при наличии разнородных выразительных средств, требующих специальной обработки. Сказанное приводит к усложнению процедур анализа и использования предметных знаний при невозможности унификации механизмов работы со знаниями в целом.

Г.М. Емельяновым, Т.В. Кречетовой и Е.П. Курашовой была предпринята попытка решить эту задачу с привлечением уровня глубинного синтаксиса ЕЯ на основе модели Семантической Эквивалентности (СЭ) с использованием грамматик деревьев (Δ-грамматик) в качестве формального аппарата математического моделирования [4]. Указанный математический аппарат, предложенный А.В. Гладким и И.А.Мельчуком [5] и расширенный разделением преобразований узлов и ветвей, позволил решить задачу моделирования синонимических преобразований ЕЯ-высказываний на уровне варьирования универсальной (абстрактной) лексикой без существенного ограничения входного ЕЯ и ПО решаемых задач. Но и данному подходу в том виде, в котором он описывается в [4], присущи серьёзные недостатки:

– на уровне глубинного синтаксиса текст представлен фразами, каждая из них соответствует простому распространенному предложению. Предложенная в [4] модель работает с совокупностями деревьев глубинного синтаксиса отдельных фраз, а к каждому дереву применимо одно или несколько правил синонимических преобразований. При этом нельзя говорить о необходимых и достаточных признаках синонимии двух фраз по анализу применимости правил и, как следствие, целесообразности трансформаций того или иного типа;

– словарная подсистема предполагается замкнутой ввиду существенной сложности описываемой словарём информации;

– отсутствует формализация компонент условий применимости правил синонимических преобразований;

– синонимические преобразования глубинных синтаксических структур в теоретическом плане проработаны не до конца. Использованный в [4] набор

правил был взят из работ Ю.Д. Апресяна [6] и И.А. Мельчука [7]. По оценке последнего, указанные правила не претендуют на полноту и возможно их расширение по результатам соответствующих исследований.

В общих чертах интерпретация результата выполнения ТЗОФ есть анализ степени близости ответа испытуемого заданному эталону, что предполагает доказательство идентичности ролей сходных понятий относительно сходных ситуаций, описываемых сравниваемыми текстами ответа испытуемого и “правильного” ответа, сформулированного разработчиком теста. Наиболее близка данной идее обработка текстов на основе коммуникативной грамматики. Хорошим примером является поисковая система “Exactus”. Тем не менее, в отличие от традиционного для поисковых систем взаимодействия “запрос – ответ”, интерпретация результатов ТЗОФ означает не столько отображение ответа на заданную ПО, сколько оценку близости нужному эталону. По оценке Г. С. Осипова [8], свойства связей между минимальными семантико-синтаксическими языковыми единицами в самой коммуникативной грамматике требуют более детального изучения. В связи с этим задача *автоматизации накопления знаний* о взаимодействии семантики, синтаксиса и морфологии ЕЯ *при установлении СЭ* текстов *является* чрезвычайно *актуальной*.

Настоящая диссертационная работа посвящена вопросам организации структуры, пополнения и использования знаний о синонимии, сформированных на основе множеств семантически эквивалентных ЕЯ-фраз, для численной оценки смысловой близости текстов. Особенность постановки задачи здесь заключается в том, что тексты, используемые для формирования знаний, вводятся пользователем без специальной подготовки в области языкознания.

Цель диссертации заключается в разработке теоретико-методологических основ организации обработки ЕЯ для задачи автоматизированного обучения и контроля знаний на основе ТЗОФ. Для достижения поставленной цели в работе решаются следующие задачи:

- анализ существующих методов моделирования семантики конструкций ЕЯ и определение общих требований, предъявляемых к механизму сравнения смыслов на функциональном уровне;
- разработка и исследование методов моделирования СЭ на уровне варьирования абстрактной лексикой;
- разработка методов накопления и систематизации знаний о морфологии и синтаксисе ЕЯ, учитывающих возможные формы выражения заданного смысла;
- моделирование и алгоритмизация механизма использования знаний о морфологии и синтаксисе языка для решения задачи кластеризации предметных и языковых знаний;
- разработка и исследование методов численной оценки смысловой близости ответа испытуемого варианту правильного ответа для интерпретации результатов теста открытой формы;
- разработка архитектуры программной системы контроля знаний, реализующей предложенные методы и модели.

Методы исследования. Для решения поставленных в работе задач были использованы методы формальной теории языков, математической логики и теории множеств, основные положения теоретической и когнитивной лингвистики, системной типологии языков и когнитологии, теории решеток и анализа формальных понятий, а также численные методы анализа данных.

Научная новизна. В диссертации разработаны теоретические основы численной оценки семантической эквивалентности текстов, их компрессии и распознавания для задач автоматизированного контроля знаний. В частности, новыми являются следующие теоретические результаты:

- теоретико-множественная модель процесса установления семантической эквивалентности для флективного языка в рамках синонимического варьирования на уровне абстрактной лексики;
- комплексный подход к решению задачи пополнения лингвистических информационных ресурсов из текстов выделением синтаксических контекстов существительных с последующим упорядочиванием знаний на основе решёток формальных понятий;
- математическая модель процесса построения формальных образов сверхфразовых единств на уровне глубинного синтаксиса, отличающаяся применением информационно-логической модели совокупности правил грамматики деревьев (Δ-грамматики) для поиска последовательности преобразований с заданными свойствами;
- математическая модель и комплекс программ формирования и кластеризации семантических отношений на основе описаний ситуаций действительности множествами эквивалентных по смыслу фраз предметно-ограниченного подмножества естественного языка;
- метод численной оценки смысловой близости текстов предметного языка для интерпретации результатов теста открытой формы;
- приближённые методы автоматизированного построения модели смыслового эталона в виде решётки формальных понятий (включая численные оценки точности получаемого решения), модель процесса интерпретации ответа испытуемого на тестовое задание открытой формы, а также метод компрессии текстовой базы знаний на основе выделенных эталонов;
- типовая архитектура программной системы контроля знаний, реализующая предложенные в работе методы и модели.

Теоретическая и практическая значимость. Основная часть диссертации носит характер теоретический характер, главы 5 и 6 наряду с теоретическими результатами описывают прикладную составляющую выполненного автором исследования. Полученные в диссертационной работе результаты, разработанные методы и комплекс программ могут быть использованы для решения широкого класса задач анализа текстов, а также сжатия информации без потери полезной смысловой составляющей. Наряду с текстами естественного языка, выделение смысловых эталонов предлагаемыми в работе методами актуально для задач распознавания и анализа семантики любых сложных информационных объектов, в том числе изображений, при формировании баз данных и зна-

ний. В частности, результаты диссертационной работы были использованы в следующих научно–исследовательских работах:

1. Грант № ТОО-3.3-408 Минобразования РФ.
2. Контракт № И 0675 ФЦП “Интеграция”, гос. рег. № 01.2.003 00918.
3. Грант РФФИ № 03-01-00055-а “Разработка математического аппарата для распознавания сверхфразовых единств в текстах”, руководитель Емельянов Г. М.
4. Грант РФФИ № 06-01-00028-а “Разработка методов автоматизированного пополнения тезауруса для задач распознавания смысловой эквивалентности текстов”, руководитель Емельянов Г. М.
5. Грант РФФИ № 10-01-00146-а “Разработка методов автоматизированного накопления и систематизации знаний о морфологии и синтаксисе естественного языка для задач семантической кластеризации текстов”, руководитель Емельянов Г. М., гос. рег. № 01201164263, 2010-2012 г.
6. ГБ НИР “Разработка и исследование математических моделей многопараметрических систем”, руководитель Емельянов Г.М., по заданию Минобрнауки РФ, гос. рег. № 01.2.007 04719, 2007-2011 г.

Достоверность теоретических результатов обеспечивается применением апробированного математического аппарата, корректностью изложения основных теоретических положений диссертационной работы с формулировкой необходимых лемм и теорем и строгостью математических доказательств. Теоретические положения диссертации иллюстрируются практическими примерами реализации компонент программной системы тестирования знаний и решения возникающих при этом инженерных задач.

Личный вклад автора. В диссертационной работе обобщены результаты, полученные лично автором или при его непосредственном участии. Постановка задачи сжатия информации (в том числе текстовой) без потери полезной смысловой составляющей принадлежит научному консультанту Емельянову Г.М., моделирования семантической эквивалентности грамматиками деревьев – Курашовой Е.П., моделирования ограниченными сетями Петри динамических информационных структур – Зайцевой Е.И. Исследования по моделированию сжатия смысловой информации на уровне глубинного синтаксиса выполнено лично автором. Решение задач пополнения лингвистических информационных ресурсов из текстов и кластеризации знаний на основе синтаксического контекста имени существительного получены как обобщение модели автоматического извлечения знаний, предложенной Степановой Н.А. применительно к генитивной конструкции русского языка и развитой совместно с автором настоящей диссертационной работы. Совместно с Корнышовым А.Н. автором развит подход к моделированию семантики конструкций ЕЯ на основе ситуаций языкового употребления. Метод численной оценки смысловой близости текстов предметного языка, а также методы автоматизированного построения модели смыслового эталона и модель процесса интерпретации ответа испытуемого на тестовое задание открытой формы (включая архитектуру программной системы контроля знаний) получены лично автором. Программы для ЭВМ разработаны

совместно с Залешиным М.В., машинные эксперименты подготовлены и проведены совместно с Емельяновым Г.М., Силановым Д.В. и Юрченко И.И.

Апробация работы. Результаты работы апробированы в докладах на конференциях, семинарах и конгрессах: 5-й, 6-й, 7-й, 8-й, 9-й Международных конференциях "Распознавание", Курск, 2001, 2003, 2005, 2008, 2010; 10-й, 12-й, 13-й, 14-й Всероссийских конференциях "Математические методы распознавания образов", Москва, 2001, 2005, Зеленогорск (Ленинградская область), 2007, Суздаль (Владимирская область), 2009; VI-й, VIII-й Всероссийских конференциях “Методы и средства обработки сложной графической информации”, Нижний Новгород, 2001, 2005; Международном семинаре Диалог’2002 “Компьютерная лингвистика и интеллектуальные технологии”, Москва, 2002; 4-й, 5-й, 6-й, 7-й, 8-й Международных конференциях “Интеллектуализация обработки информации”, Алушта (Автономная Республика Крым, Украина), 2002, 2004, 2006, 2008, Пафос (Республика Кипр), 2010; 6-й, 7-й, 8-й, 9-й, 10-й Международных конференциях “Распознавание образов и анализ изображений: новые информационные технологии”, Великий Новгород, 2002, Санкт-Петербург, 2004, 2010, Йошкар-Ола, 2007, Нижний Новгород, 2008; VI-м Международном конгрессе по математическому моделированию, Нижний Новгород, 2004; XVIII Международной научно-методической конференции “Математика в вузе”, Санкт-Петербург, 2005; 6-й, 7-й, 8-й Международных научно-технических конференциях “Интерактивные системы: проблемы человеко-компьютерного взаимодействия”, Ульяновск, 2005, 2007, 2009; XIII-й, XIV-й, XV-й, XVI-й, XVII-й, XVIII-й научных конференциях преподавателей, аспирантов и студентов НовГУ “Дни науки в НовГУ”, Великий Новгород, 2006, 2007, 2008, 2009, 2010; юбилейной научно-практической конференции “Великий Новгород – город университетский”, Великий Новгород, 2003; научных семинарах кафедр “Программного обеспечения вычислительной техники и автоматизированных систем” и “Информационных технологий и систем” Новгородского государственного университета имени Ярослава Мудрого с 2001 по 2011 годы.

Публикации. Всего по теме диссертации опубликовано 70 работ, среди них одна монография, 2 – учебники и учебно-методические пособия, 16 статей в журналах, входящих в перечень, рекомендованный ВАК для публикации основных результатов докторских диссертаций. Имеется свидетельство о регистрации программы для ЭВМ. В трудах международных конференций представлено 27 работ, в трудах всероссийских – 6 работ. Перечисленные работы достаточно полно отражают содержание диссертации.

Структура и объем диссертации. Диссертация состоит из введения, шести глав, заключения, приложения и списка литературы. Общий объем диссертации составляет 318 страниц машинописного текста. Основная часть работы изложена на 226 страницах и содержит 73 рисунка и 13 таблиц. Список литературы включает 170 наименований.

На защиту выносятся следующие основные положения:

1. Комплексная методика пополнения лингвистических информационных ресурсов из текстов и выделения классов смысловой эквивалентности на основе решёток формальных понятий.

2. Формальная концептуальная модель сжатия смысловой информации на основе классов смысловой эквивалентности для уровня абстрактной лексики.
3. Математическая модель и комплекс программ формирования и кластеризации семантических отношений в виде классов формальных понятий решётки и основанный на указанной модели метод выделения смыслового эталона на множестве эквивалентных по смыслу фраз предметно-ограниченного подмножества естественного языка.
4. Модель процесса интерпретации ответа испытуемого на тестовое задание открытой формы распознаванием смысловых эталонов, а также метод компрессии текстовой базы знаний с применением указанных эталонов.
5. Метод численной оценки смысловой близости текстов предметного языка для интерпретации результатов теста открытой формы.

Диссертация посвящена разработке, исследованию и обоснованию методов математического моделирования СЭ в естественном языке, вопросам автоматической генерации моделей указанного явления, используемых для решения научных и прикладных задач обработки смысловой информации численными методами; созданию комплекса программ, реализующих предложенные методы, что отвечает паспорту специальности 05.13.18 – “Математическое моделирование, численные методы и комплексы программ”.

КРАТКОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Содержание работы. Во введении обоснована актуальность темы работы, дан краткий аналитический обзор современного состояния проблематики и литературы по теме исследования, сформулированы цели и задачи, определена структура диссертации.

Первая глава посвящена общей постановке задачи сравнения текстов на предмет СЭ. Вводится понятие конструкции ЕЯ – последовательности знаков в некоторой знаковой системе, которая может быть использована для фиксации некоторого числа высказываний этого ЕЯ в памяти ЭВМ. Для решения задачи установления СЭ в качестве модели семантики конструкций ЕЯ предложено использовать модель Ситуации Языкового Употребления (СЯУ).

Определение 1. Под СЯУ (ситуацией употребления ЕЯ) понимается описание нового социального опыта (содержания совместных действий) средствами этого ЕЯ. Данное описание выполняется в некоторой знаковой системе с целью обобщения и передачи знаний от человека к человеку.

Фиксируемый ситуацией S языковой контекст представляется тройкой:

$$S = \left(O, R, T^S\right), \qquad (1)$$

где O есть множество объектов-участников S; R – множество отношений между $o \in O$; T^S – множество форм языкового описания S. Модель (1) предложена автором в работе [6] из “Списка основных публикаций автора по теме диссертации” (далее – “Списка публикаций автора”) совместно с А.Н. Корнышовым как основа концептуально-ситуационного моделирования ЕЯ-высказываний.

Сама задача СЭ при этом формулируется следующим образом.

Задача 1. Дано множество ЕЯ-текстов *G*. Требуется: по результатам разбора каждого $T_i \in G$ выявить:

– множество $V(T_i)$ ситуаций, описываемых T_i;

– множество $M(T_i)$ объектов и/или понятий, значимых в ситуациях из множества $V(T_i)$;

– тернарное отношение $I \subseteq G \times M \times V$, ставящее в соответствие каждому $m \in M : M = \bigcup_i M(T_i)$ ту ситуацию $v \in V : V = \bigcup_i V(T_i)$, в которой он фигурирует относительно T_i. Далее на основе отношения I необходимо выделить группы текстов, сходных по встречаемости объектов в одних и тех же ситуациях.

Далее в главе анализируются достоинства и недостатки модели СЭ на базе подхода "Смысл⇔Текст". На основе теории Анализа Формальных Понятий (АФП) автором предлагается комплексный подход к решению задачи пополнения лингвистических информационных ресурсов из текстов с выделением классов смысловой эквивалентности как классов Формальных Понятий (ФП) в решётке. Данный подход естественно согласуется с *Задачей 1*. Сами тексты при этом объединяются в классы по сходству признаков сочетаемости слов относительно языковых контекстов СЯУ. Основные результаты главы представлены в работах [1, 6, 7, 29, 32, 36, 38, 41, 50, 58] из списка публикаций автора.

Вторая глава посвящена решению проблемы полноты представления смысла при описании синонимического варьирования на уровне глубинного синтаксиса. В соответствии с поставленной в первой главе задачей иерархизации знаний о синонимии, строится формальная концептуальная модель сжатия смысловой информации на основе классов СЭ, выделяемых для абстрактной лексики. При этом содержательную основу сжатия информации составляют формальные образы сверхфразовых единств на уровне глубинного синтаксиса. Для теоретического обоснования алгоритмической разрешимости процесса построения последних вводится информационно-логическая модель совокупности правил Δ-грамматики на основе аппарата ограниченных сетей Петри.

Пусть Π^R – множество правил синонимических преобразований деревьев глубинного синтаксиса в рамках стандартных Лексических Функций (ЛФ), а $r(\pi)$ – условие применимости правила $\pi \in \Pi^R$. Содержательно $r(\pi)$ есть совокупность требований к лексическим единицам, заменяемым посредством π. При этом $r(\pi)$ выступает в роли прецедента как типичного представителя таксона π, с которым отождествляется класс СЭ на уровне абстрактной лексики.

Определение 2. Лексической Синонимической Конструкцией (ЛСК) будем далее называть комплекс лексических единиц и связывающих их глубинно-синтаксических отношений, замена которого описывается некоторым $\pi \in \Pi^R$. Каждой ЛСК соответствует свое ключевое слово C_0, либо непосредственно входящее в нее, либо выраженное в значениях ЛФ от C_0 в комплексе составляющих ЛСК лексических единиц.

Представим вход правила $\pi_i \in \Pi^R$ как описание поддерева, заменяемого правилом. Тогда определение возможности применения преобразований из Π^R к заданному дереву есть анализ применимости каждого $\pi_i \in \Pi^R$, с выделением ключевого слова ЛСК и представлением результата в виде списка пар:

$$\left\{(\pi_i, C_0(i)) : i = 1,\ldots,\left|\Pi^R\right|\right\}. \quad (2)$$

В работе некоторого правила $\pi \in \Pi^R$ в общем случае следует выделить два состояния: соответствующее заменяемому дереву T_1^π и соответствующее заменяющему дереву T_2^π. Иными словами, мы имеем простейший случай задачи достижимости ЛСК с заданными свойствами на информационном пространстве, заданном входами и выходами правил $\pi \in \Pi^R$. Условие $r(\pi)$ представляет собой формальное описание допустимости перехода из состояния T_1^π в T_2^π.

В общем случае для каждого $\pi \in \Pi^R$ следует рассматривать множество R_π условий применимости, из которых для срабатывания правила должно выполниться минимум одно $r_i(\pi) \in R_\pi$. Правило π может быть применено к дереву T_1^π, если выполняется некоторое $r_j(\pi) \in R_\pi : \vee_{j=1}^m r_j(\pi) = true$, где $m = \left|R_\pi\right|$. Обозначим $\vee_{j=1}^m r_j(\pi)$ далее как r_{12}. Условие r_{12} следует интерпретировать как "определение события, разрешающего переход от T_1^π к T_2^π". Применение правила $\pi \in \Pi^R$ сводится к выполнению перехода:

$$\pi(r_{12}) : T_1^\pi \xrightarrow{\pi(r_{12})} T_2^\pi. \quad (3)$$

Отдельному правилу соответствует элементарная сеть Петри вида

$$N = \{P, T, F, H, M_0\}. \quad (4)$$

При этом множество состояний правила есть множество P позиций сети: $P = \{p_1, p_2\}$, где $p_1 \Leftrightarrow T_1^\pi$, а $p_2 \Leftrightarrow T_2^\pi$. Множество возможных переходов T представлено единственным переходом из T_1^π в T_2^π: $t = \pi(r_{12}) : p_1 \xrightarrow{t} p_2$. Компоненты F и H есть отображения $F : P \times T \to \{0,1\}$ и $H : T \times P \to \{0,1\}$, соответственно. Для сети (4) $F(p_1, t) = 1$, $F(p_2, t) = 0$, $H(t, p_1) = 0$, $H(t, p_2) = 1$, а число допустимых разметок сети, отождествляемых со сценариями, здесь равно двум. Начальной маркировке соответствует вектор $M_0 = (1,0)$, второй из допустимых маркировок – вектор $M = (0,1)$.

Множество правил Δ-грамматики, представленных элементарными сетями Петри, можно рассматривать как множество исходных объектов-примитивов для построения в терминах ограниченных сетей Петри модели системы правил

некоторого подмножества множества Π^R рассматриваемой Δ-грамматики с определением структурных взаимосвязей между ними. При этом сама система формируется следующим образом: для каждой пары правил $\{\pi_1, \pi_2\} \subset \Pi^R$, $\pi_1 \neq \pi_2$, входящих в систему, обязательным является выполнение следующего условия: либо вход правила π_2 является выходом для π_1, либо наоборот, вход у π_1 является выходом для π_2.

Пусть N_i – сеть Петри, построенная из примитивов, каждый из которых моделирует работу правила из некоторого подмножества правил заданной Δ-грамматики, образующих систему. Состоянию системы соответствует активизация входа/выхода некоторого правила.

Теорема 1. Сеть N_i является безопасной в течение всего времени функционирования системы.

Доказательство следует из наложенного на структуру сети ограничения относительно числа позиций, инцидентных переходу.

Последовательность применяемых правил моделируется последовательностью $\tau = \left(t_i^1, t_i^2, \ldots, t_i^k\right)$ срабатываний переходов:

$$T_1^\pi \xrightarrow{\pi_1(r_{12})} T_2^\pi \xrightarrow{\pi_2(r_{23})} T_3^\pi \to \ldots \to T_k^\pi \xrightarrow{\pi_k\left(r_{k\,k+1}\right)} T_{k+1}^\pi, \qquad (5)$$

где $t_i^1 \Leftrightarrow \pi_1(r_{12})$, $t_i^2 \Leftrightarrow \pi_2(r_{23})$, … , $t_i^k \Leftrightarrow \pi_k\left(r_{k\,k+1}\right)$. При этом происходит последовательная смена разметок:

$$M_{0i} \xrightarrow{t_i^1} M_i^1 \xrightarrow{t_i^2} M_i^2 \to \ldots \to M_i^{k-1} \xrightarrow{t_i^k} M_i^k, \qquad (6)$$

где $M_{0i} \Leftrightarrow T_1^\pi$, $M_i^1 \Leftrightarrow T_2^\pi$, … , $M_i^{k-1} \Leftrightarrow T_k^\pi$, $M_i^k \Leftrightarrow T_{k+1}^\pi$.

При этом множество достижимости $R(N_i)$ сети N_i находится в зависимости от задания начальной разметки M_{0i}. Функционирование системы описывается в терминах последовательностей срабатываний переходов $t_i^1, t_i^2, \ldots, t_i^{k-1}, t_i^k$, каждая из которых есть слово τ в языке $L(N_i)$.

Задача приведения деревьев T_1^π и T_{k+1}^π к виду с одинаковой ЛСК фактически включает в себя три задачи:

1) определение достижимости разметки M_i^k из начальной разметки M_{0i}. Данная задача есть поиск слова $\tau \in T_i^* \left| M_{0i} \xrightarrow{\tau} M_i^k \right.$, где T_i^* – множество всех слов в алфавите T_i;

2) задача обратимости слова τ: если $\tau \in T_i^* \left| M_{0i} \xrightarrow{\tau} M_i^k \right.$, то существует ли слово $\tau' = \left(t_i^{k'}, t_i^{(k-1)'}, \ldots, t_i^{2'}, t_i^{1'}\right)$:

$$M_{0i} \xleftarrow{t_i^{1'}} M_i^1 \xleftarrow{t_i^{2'}} M_i^2 \leftarrow \ldots \leftarrow M_i^{k-1} \xleftarrow{t_i^{k'}} M_i^k, \qquad (7)$$

где $M_{0i} \Leftrightarrow T_1^{\pi}$, $M_i^1 \Leftrightarrow T_2^{\pi}$, … , $M_i^k \Leftrightarrow T_{k+1}^{\pi}$;

3) задача определения оптимального слова $\tau \in T_i^* \left| M_{0i} \xrightarrow{\tau} M_i^k \right.$. Суть: если существуют несколько слов $\tau_1, \tau_2, \ldots, \tau_l$, описывающих последовательности смены разметок $M_{0i} \xrightarrow{\tau_1} M_i^k$, $M_{0i} \xrightarrow{\tau_2} M_i^k$, … , $M_{0i} \xrightarrow{\tau_l} M_i^k$, то в качестве оптимального берется обратимое слово минимальной длины.

Для решения указанных задач проводится исследование языка $L(N_i)$. Как результат исследования сформулированы лемма и ряд теорем.

Лемма 1. Проблема достижимости заданной разметки M_i^k из начальной M_{0i} в сети N_i разрешима.

Теорема 2. Все символы-переходы $t_i^j \in T_i$ сети N_i различны.

Теорема 3. Проблема определения обратимости слова $\tau \in T_i^* \left| M_{0i} \xrightarrow{\tau} M_i^k \right.$ языка $L(N_i)$ разрешима.

Теорема 4. Проблема определения оптимального слова $\tau \in T_i^* \left| M_{0i} \xrightarrow{\tau} M_i^k \right.$ (T_i^* – множество всех слов в алфавите T_i) в языке $L(N_i)$ является разрешимой.

С целью сокращения перебора при поиске оптимального слова вводится описание состояния системы правил парой активизированных информационных элементов и определенное как сценарий. Введение в описание сценария “ссылки назад” для обратного просмотра по дереву сценариев позволяет сократить перебор при поиске оптимального слова. При этом путь к целевому состоянию строится на основе заданного начального сценария и массивов ссылок на описания входных/выходных деревьев и условий применимости правил.

На основе предложенного исчисления сценариев на информационном пространстве системы правил Δ-грамматики проведен анализ событий, вызывающих активизацию входов/выходов правил как элементов информационного пространства. Как результат предложена и исследована функционально-логическая модель входа/выхода правила в виде ограниченной сети Петри:

$$N_{\pi(k)} = \left\{ P_{\pi(k)}, T_{\pi(k)}, F_{\pi(k)}, H_{\pi(k)}, C, M_{0\pi(k)} \right\}, \qquad (8)$$

где множество позиций $P_{\pi(k)}$ соответствует множеству состояний информационного элемента, а каждое состояние отождествляется с очередным пройденным узлом $w_i^{\pi} \in W_k^{\pi}$; каждому из переходов $t_{\pi(k)}^i \in T_{\pi(k)}$ соответствует совокупность требований лексической, грамматической части и метки входящей ветви узла w_i^{π}; $F_{\pi(k)}$ и $H_{\pi(k)}$ есть матрицы инцидентности, аналогичные соответствующим компонентам структуры (4); $C = \{c_1, c_2, c_3, c_4, c_5\}$ – множество цветов маркера; $M_{0\pi(k)}$ – начальная разметка. Каждому цвету соответствует свой способ использования информационного элемента: c_1 – анализ применимости пра-

вила, c_2 – синтез дерева на выходе правила, c_3 – определение ключевого слова ЛСК, c_4 – расстановка композиционных меток в анализируемом дереве. Во избежание развертывания бесконечных процессов в сети множество C содержит нейтральный маркер c_5, запрещающий срабатывание перехода.

Сеть $N_{\pi(k)}$ обладает рядом свойств, позволяющих оценить адекватность порождаемых ей процессов моделируемым процессам, порождаемым входом/выходом правила π при анализе его применимости к некоторому дереву либо синтезе результирующего дерева по шаблону T_k^{π}.

Теорема 5. Все порождаемые сетью $N_{\pi(k)}$ процессы конечны.

Теорема 6. Сеть $N_{\pi(k)}$ является ограниченной.

Далее в главе исследуется алгоритмическая сложность построения суммарного образа для деревьев глубинного синтаксиса с одинаковой ЛСК.

Теорема 7. Задача установления функционального соответствия деревьев $T_{\chi 1}$ и $T_{\chi 2}$ принадлежит классу P комбинаторных задач с временной оценкой n^D, где $T_{\chi 1} = \langle W_{\chi 1}, V_{\chi 1} \rangle$, $T_{\chi 2} = \langle W_{\chi 2}, V_{\chi 2} \rangle$, $W_{\chi 1}$ и $W_{\chi 2}$ – множества узлов, $V_{\chi 1}$ и $V_{\chi 2}$ – множества ветвей, $n = \max(|W_{\chi 1}|, |W_{\chi 2}|)$, $D = \sum_{i=1}^{|V^R|} \varphi(a_i)$, φ есть матрица ограничений на характер ветвления в дереве, V^R – словарь пометок на ветвях.

Само функциональное соответствие определяется следующим образом.

Определение 3. Деревья T_1 и T_1' считаются изоморфными с точностью до функционального соответствия, если в дереве T_1' из узла α_{11}' в узел α_{12}' идет ветвь с некоторой пометкой тогда и только тогда, когда в дереве T_1 из узла α_{11} в узел α_{12} идет ветвь с той же пометкой. При этом узел α_{11}' должен отвечать требованиям, содержащимся в узле α_{11}, а узел α_{12}', соответственно, требованиям, содержащимся в узле α_{12}. В таком случае считается, что узел α_{11}' функционально соответствует узлу α_{11}, а узел α_{12}' – узлу α_{12}.

Завершает главу качественный анализ взаимодействия процессов построения сверхфразовых единств в текстах и установления их СЭ. Рассматривается использование для этой цели списков вида (2). Приводится пример построения формального образа сверхфразового единства для четырех простых распространенных предложений русского языка. Основные результаты главы представлены в работах [1–5,8–10,18–24,45,46,52–55] из списка публикаций автора.

В третьей главе диссертации решается задача формирования и классификации семантических отношений как основы знаний о синонимии. При этом основополагающим является понятие прецедента класса СЭ.

За основу формального описания прецедента в главе берётся введённое Б.Х. Парти и В.Б. Борщевым представление Лексического Значения (ЛЗ) слова посредством теории – совокупности аксиом (meaning postulates, [9]), каждая из ко-

торых описывает отдельное свойство обозначаемого словом объекта реального мира. Сама теория ЛЗ слова w_i, заменяемого посредством некоторого правила $\pi \in \Pi^R$, описывается посредством структуры:

$$Lm(w_i) = \left(w_i, L^M\right), \tag{9}$$

где L^M – список структур, задающих отношения между словами и понятиями. При этом отдельный элемент списка L^M может представлять как бинарное отношение между парой понятий C_1 и C_2:

$$M_p = (R_2, C_1, C_2), \tag{10}$$

так и рекурсивно определяемое отношение произвольной арности вида

$$M'_p = \left(R_n, C, L^M\right) \tag{11}$$

либо

$$M''_p = \left(R_c, L^M\right), \tag{12}$$

где $R_c \in \{\vee, \&, \neg\}$. Посредством L^M в (11) задается связь понятия C с другими словами и понятиями.

Утверждение 1. Если имеется описание теории $Lm(w_i) = \left(w_i, L^M\right)$ ЛЗ слова w_i структурой (9), то смысл слова определяется набором характеристических функций (ХФ) ChF_{hi} таких, что выполняются следующие условия:

1. В списке L^M содержится структура $M_p = (R_2, C_1, C_2)$ вида (10) (обозначим ее как ChF_{Val}), при этом $ChF_{hi}(w_i) = C_2$, где C_2 – некоторое известное понятие, а L^M может быть третьим аргументом структуры (11).

2. Существует структура (далее обозначаемая как ChF_{Name}) либо вида (10), и при этом $M_p = \left(ChF_{hi}, C_1^1, C_2^1\right)$, либо вида (11), и $M'_p = \left(ChF_{hi}, C, L^M\right)$, но в обоих случаях ChF_{hi} – имя известного смыслового отношения.

3. Если ChF_{Name} есть первая структура, удовлетворяющая условию (2) при обратном просмотре списка L^M от ChF_{Val}, и $L^{M'} \subset L^M$ есть список, такой, что либо $L^{M'} = \left\{\left(ChF_{hi}, C_1^1, C_2^1\right), \ldots, ChF_{Val}\right\}$, либо $L^{M'} = \left\{\left(ChF_{hi}, C, L^M\right) : ChF_{Val} \in L^M\right\}$, то каждое последующее утверждение в $L^{M'}$ должно иметь как минимум один общий аргумент, являющийся обозначением некоторой переменной, с предыдущим утверждением.

Опираясь на понятия экстенсионала и интенсионала из интенсиональной логики Ричарда Монтегю, в главе решается задача обобщения знаний в виде структур (9) с применением АФП.

Лексическое значение слова, описываемое формализованной теорией (9), есть денотация. В логике ей ставится в соответствие экстенсионал как класс сущностей, которые определяются посредством структуры (9). При этом внешне различные описания теорий одного и того же ЛЗ определяют единое множество ХФ,

задаваемых согласно *утверждению 1*. Характеристические функции (в том числе определяемые рекурсивно посредством структур (11) и (12)) задают набор формальных признаков для элементов толкования лексического значения. Эти признаки и определяют интенсионал обобщенной теории заданного ЛЗ. Ключевое правило обобщения утверждений независимых вариантов теории ЛЗ определяется введением в рассмотрение области, которую образуют элементы толкования заданного ЛЗ в решетке формальных понятий. При этом вычислительная сложность процесса обобщения теорий заданного ЛЗ зависит исключительно от мощности множества характеристических функций.

Для автоматизации получения знаний, представляемых структурами (9)–(12), в **разделе 3.5** решается задача формирования множества отношений R в модели (1) на основе множеств СЭ-фраз. Отношения в рамках структур (9)–(12) будут составлять подмножество множества R согласно *определению 1*.

Рассмотрим текст $T_i \in T^S$ с точки зрения символов, которые его составляют. У каждого $T_i \in T^S$ в модели (1) выделяется некоторую неизменную часть T_i^C, общая для всех $T_i \in T^S$, и флективная часть T_i^F. На множестве T_i^F выражаются синтагматические зависимости, которые задаются с помощью синтаксических отношений и определяют возможность сосуществования словоформ в линейном ряду. Аналогично $W_{ij} = W_{ij}^C \cup W_{ij}^F$, где W_{ij} – буквенный состав слова, $W_{ij}^C \subset T_i^C$ – неизменная, $W_{ij}^F \subset T_i^F$ – флективная часть. Для формирования искомого множества R попарным сравнением W_{ij} различных T_i требуется найти:

1) W_{ij}^C и W_{ij}^F каждого W_{ij} при $\left|W_{ij}^C\right| \to \max$;

2) отношение R_q, определяющее допустимость сочетания $\left(W_{ij}^F, W_{ik}^F\right)$, $k \neq j$.

Введём в рассмотрение индексное множество J для неизменных частей всех слов, употребленных во всех фразах из T. Тогда упорядоченная совокупность индексов $j \in J$ неизменных частей слов, присутствующих в $T_i \in T^S$, будет моделью линейной структуры этой фразы (далее обозначается как $L(T_i)$). Для построения множества R в составе структуры (1) необходимо найти совокупность указанных моделей, удовлетворяющих требованиям проективности.

Пусть $h(j, L(T_i))$ – позиция индекса j в модели $L(T_i)$. Тогда множество связей для $L(T_i)$ определяется как $D: T_i \to \{(h(j, L(T_i)), h(k, L(T_i))) : j \neq k\}$.

Определение 4. Связь $d_{qi} = (h(j, L(T_i)), h(k, L(T_i)))$ является *допустимой для модели* $L(T_i)$, если $\exists\{T_l, T_m\} \subset T$, $l \neq m$, причем и $L(T_l)$, и $L(T_m)$ содержат в качестве подпоследовательности либо $\{j,k\}$, либо $\{k, j\}$. При этом пара индексов (j,k) соответствует одной синтагме, а индекс q – типу синтаксического отношения, которое ей соответствует.

Положим, что для $\forall T_i \in T$, $i = 1,\ldots,\left|T^S\right|$, все $d_{qi} \in D(T_i)$ удовлетворяют *определению 4*.

Определение 5. Будем считать, что модель $L(T_i)$ проективна относительно множества R в (1), если $\sum_{q=1}^{|D(T_i)|} \Delta_{qi} \le |L(T_i)|$, где $\Delta_{qi} = |h(j, L(T_i)) - h(k, L(T_i))|$.

На основе $\bigcup_i D(T_i)$ формируется граф синтагм $\left(V^J, I^J\right)$. Элементами множества вершин V^J являются множества пар (j,k), $\{j,k\} \subset J$, сгруппированных по некоторому индексу k. Множества E_1 и E_2, входящие в V^J, будут соединены ребром из I^J, если $\exists\{j,k,m\} \subset J$: $(j,k) \in E_1$, $(k,m) \in E_2$ и $j \ne m$. Анализом $\left(V^J, I^J\right)$ строится дерево-прецедент $\left(V_1^J, I_1^J\right)$ для $\bigcup_i T_i$, $i = 1,\ldots,\left|T^S\right|$. Формально

$$V_1^J = J,\ I_1^J = \left\{(j,k) : \exists E \in V^J, (j,k) \in E\right\}. \qquad (13)$$

При этом $k \in V_1^J$ соответствует корню дерева $\left(V_1^J, I_1^J\right)$, если $\exists E_1 \in V^J$, в котором пары индексов сгруппированы по k, $|E_1| > 1$, а k не содержится ни в одной паре индексов для $\forall E_2 \in V^J$: $E_1 \ne E_2$. Поскольку наибольший интерес для формирования множества R представляют ситуации (1) с двумя и более участниками, то число дочерних узлов у корня полагается больше одного.

Использованием маршрутов в дереве (13) задача выделения классов отношений множества R в модели (1) естественно решается методами АФП. При этом множество флексий рассматривается как множество формальных объектов $G^F = \left\{f_{ij} : f_{ij} = \bullet\left(W_{ij}^F\right)\right\}$, где $i = 1,\ldots,\left|T^S\right|$, а символом "$\bullet$" обозначается операция конкатенации, которая последовательно выполняется над символами из W_{ij}^F.

Введем в рассмотрение Формальный Контекст (ФК):

$$K^F = \left(G^F, M^F, I^F\right), \qquad (14)$$

где $M^F = G^F$, а $I^F \subseteq G^F \times M^F$. При этом $I^F = \left\{\left(f_{ij}, f_{ik}\right) : s(j,k) = true, \{j,k\} \subset J\right\}$.

Отношение s определяется рекурсивно на основе $\left(V^J, I^J\right)$:

1) $s(j_1, j_1) = true$;
2) $s(j_1, j_2) = true$ в одном из следующих двух случаев:
   – $\exists E_1 \in V^J$: $(j_1, j_2) \in E_1$, причем $\exists j_3 \in J$, для которого $s(j_2, j_3) = true$;
   – $\exists (E_1, E_2) \in I^J$: $\exists j_3 \in J$, при этом $(j_1, j_3) \in E_1$, $(j_3, j_2) \in E_2$, а $s(j_3, j_2) = true$.

Модель (14) выделяет классы в R по характеру изменения флективной части зависимого слова в каждом из $R_q \in R$ с учетом бинарности последнего. Далее в главе рассматривается задача поиска флексий для слов в составе Расщеп-

ленного Предикатного Значения (РПЗ) как совокупности вспомогательного глагола (связки) и некоторого существительного, называющего ситуацию.

Пусть $T_i^{Cnc} = \{w_{ij} : w_{ij} = \bullet(W_{ij})\}$. Положим также, что $\exists T_i^P \subset T_i$, определяющее последовательность $P_i^{Cnc} = \{u_k : u_k = \bullet(W_k^P), \bigcup_k W_k^P = T_i^P\}$, где $W_k^P \in T_i$ – последовательность символов слова, для которого не выделены неизменная и флективная часть.

Лемма 2. Последовательность P_i^{Cnc} содержит предикатное слово (глагол или слово, производное от него), если $\exists\{j,0,k\} \subset L(T_i)$: $\{w_{ij}, u_1, \ldots, u_p, w_{ik}\} \subset T_i^{Cnc}$, где $\{u_1, \ldots, u_p\} = P_i^{Cnc}$, $p = \left|P_i^{Cnc}\right|$.

Пусть для последовательности P_i^{Cnc} выполняется условие *леммы 2*.

Лемма 3. Слово $u_k \in P_i^{Cnc}$ принадлежит РПЗ, если $\exists T_j \in T^S$: $L(T_j) \neq L(T_i)$, а $u_k \in P_j^{Cnc}$, где P_j^{Cnc} также отвечает условию *леммы 2*. При этом $\neg\exists T_k \in T^S$: $P_k^{Cnc} \subset P_i^{Cnc}$, а $L(T_k) \neq L(T_j)$ и $L(T_k) \neq L(T_i)$.

Замечание. При выполнении условия *леммы 3* u_k может быть зависимым словом в составе РПЗ.

Пусть $P_i^{Cnc'}$ – последовательность слов, удовлетворяющих *лемме 3*.

Теорема 8. Для формирования структуры (14) при наличии РПЗ либо конверсива необходимо и достаточно найти множество $T' \subset T^S$: $T' = \left\{T_i : \left|P_i^{Cnc'}\right| \to \max\right\}$.

Для $\forall u_k \in \bigcup_i P_i^{Cnc'}$, $T_i \in T'$, неизменная и флективная часть выделяются сравнением буквенного состава со всеми $u_j \in \bigcup_l P_l^{Cnc}$: $T_l \in (T^S \setminus T')$. При этом необходимо, чтобы $2\left|W_k^C\right| > \left|W_k^F\right| + \left|W_j^F\right|$, где $W_k^P = W_k^C \cup W_k^F$, а $W_j^P = W_j^C \cup W_j^F$.

Замечание. Если $P_i^{Cnc'} \cap P_i^{Cnc} \neq \varnothing$, то $\forall u_m \in (P_i^{Cnc} \setminus P_i^{Cnc'})$ есть предлог и представляется вместе со словом слева от него в последовательности P_i^{Cnc}.

С учетом $P_i^{Cnc'}$ дерево (13) преобразуется следующим образом:

1) корень изменяется с $k = 0$ на значение k для $u_k \in P_i^{Cnc'}$ с максимальной встречаемостью в разных T_i^{Cnc} относительно заданной СЯУ;
2) левое поддерево остается без изменений;
3) правое поддерево перевешивается на узел j для $u_j \in P_i^{Cnc'}$ наименьшей встречаемости;
4) в паре $\{u_l, u_m\} \subset P_i^{Cnc'}$ дочерним будет узел слова с меньшей встречаемостью.

В итоге основу формирования модели (14) составляют те T_i, которые наиболее полно представляют языковой контекст заданной ситуации (1).

Основные результаты главы отражены в работах [1, 11–15, 25, 27, 28, 30–35, 47–49, 51, 53, 55–57, 61, 63] из списка публикаций автора.

Четвертая глава посвящена моделированию семантики синтаксического контекста существительного как основы формирования отношений в структурах (9)– (12). Такой контекст есть последовательность соподчинённых слов:

$$S_{ki} = \{v_1, \ldots, v_{n(k,i)}, m_{ki}\}, \qquad (15)$$

где v_1 – предикатное слово; m_{ki} – существительное, обозначает некоторое понятие, значимое в ситуации v_1; $\forall v_l \in \{v_2, \ldots, v_{n(k,i)}\}$ – некоторое существительное; k – номер последовательности среди выявленных из текста T_i; $n(k,i)$ – число соподчиненных существительных последовательности.

Утверждение 2. При одновременном наличии $S_{ki} = \{v_1, \ldots, v_{n(k,i)}, m_{ki}\}$ и $S_{1ki} = \{v_1, m_{ki}\}$ в разных текстах множества T^S структуры (1) для заданной СЯУ имеет место частичная СЭ (относительно m_{ki}).

Утверждение 3. При $R_q(v_1, v_2) = true$ возможно установление указанного отношения между v_1 и любым словом последовательности (15) вне зависимости от существующих отношений.

Замечание. На основании *утверждения 3* справедливо будет утверждать, что $\forall v_l \in \{v_2, \ldots, v_{n(k,i)}\}$ в составе последовательности (15) обозначает некоторое понятие, значимое в ситуации v_1, наравне с m_{ki}. Таким образом, если $V(T_i)$ есть множество ситуаций, описываемых текстом T_i, а $M(T_i)$ есть соответствующее ему множество объектов согласно постановке *задачи 1*, то для любой S_{ki} $\{v_2, \ldots, v_{n(k,i)}, m_{ki}\} \subset M(T_i)$. Причем $V(T_i) = \bigcup_k (S_{ki} \setminus \{m_{ki}\})$.

Далее в главе рассматривается задача концептуальной кластеризации текстов методами АФП на основе последовательностей (15), выделяемых по результатам синтаксического разбора отдельных предложений. Описываются алгоритмы формирования множеств $M(T_i)$ и $V(T_i)$, а также отношения I на основе синтаксического разбора исходных текстов множества G согласно постановке *задачи 1*, а также выделение конверсивов и РПЗ на основе моделей (15). Разбор осуществлялся с помощью программы “Cognitive Dwarf” (ООО “Когнитивные технологии”), показавшей при тестировании самые точные результаты.

Основные результаты главы представлены в работах [1, 16, 37, 39, 40, 59, 62] из списка публикаций автора.

В пятой главе модель (1) расширяется введением произвольных отношений между объектами $o \in O$ в множество R. Рассмотренное в **разделе 3.5** индексное множество J определяется для неизменных частей всех слов, употреб-

ленных в более чем одной фразе из T^S. При этом удвоенная длина общей неизменной части пары слов всегда больше суммы длин флективных частей.

Пусть L^S есть множество моделей линейных структур фраз из T^S на J.

Лемма 4. Пара индексов $\{j_1, j_2\} \subset J$ соответствует словам-синонимам, если $\exists\{L(T_1), L(T_2)\} \subseteq L^S : L(T_1) = J_1 \bullet \{j_1\} \bullet J_2$ и $L(T_2) = J_1 \bullet \{j_2\} \bullet J_2$, где $J_1 \subset J$, $J_2 \subset J$, а "$\bullet$" есть операция типа конкатенации над множеством *J*.

Пусть P^J – множество пар, отвечающих условию *леммы 4*. Заменим индексы, вошедшие в пары из P^J, на некоторые $j \in (\mathrm{N} \setminus J)$ во всех $L \in L^S$. Обозначим преобразованное L^S как $L^{S'}$, множество заменяемых индексов – как J^P, а множество индексов, на которые идёт замена, – как $J^{P'}$, $J^{P'} \cap J^P = \varnothing$. Фактически каждая модель в $L^{S'}$ задается на множестве $\left(J \setminus J^P\right) \cup J^{P'}$.

Теорема 9. Индексы с максимальной встречаемостью в разных моделях из $L^{S'}$ соответствуют существительным, обозначающим участников ситуации (1).

Пусть J^N есть множество индексов, удовлетворяющих *теореме 9*, $L_1(T_i) \in L^{S'}$, а $L_2(T_i)$ – модель линейной структуры предложения T_i, но относительно J^N. Обозначим множество моделей второго вида как L^N. Положим также, что имеется $L_j^{S'} \subset L^{S'}$ такое, что для всех $L_1(T_i) \in L_j^{S'}$ модели $L_2(T_i)$ одинаковы и соответствуют некоторой $L_2(T_j) \in L^N$, $T_j \in T^S$.

Теорема 10. Индексы $j \notin J^N$ с максимальной встречаемостью в различных моделях $L_1(T_i) \in L_j^{S'}$ соответствуют либо словам-наречиям, либо прилагательным, либо опорным существительным в составе генитивных конструкций.

Обозначим множество индексов, удовлетворяющих *теореме 10*, как J^A. Установление синтаксических ролей и выделение флексий для слов с индексами из $\left(\left(J \setminus J^P\right) \cup J^{P'}\right) \setminus \left(J^N \cup J^A\right) \cup \{0\}$ производится аналогично выявлению указанной информации у слов в составе РПЗ описанным в разделе 3.5 способом. При этом вместо индексов с ненулевым значением рассматриваются индексы из $J^N \cup J^A$.

Множество R в расширенной указанным образом модели (1) отражает возможные виды сочетаний главных и зависимых слов по основам и флексиям с учетом возможного неприсутствия слова во всех фразах множества T^S.

Для *численной оценки* схожести СЯУ на основе модели (1) в **разделе 5.2** вводится представление последней в виде формального контекста:

$$K^S = \left(G^S, M^S, I^S\right), \tag{16}$$

где множество G^S составляют основы слов, входящих во фразы из множества T^S и зависимые по отношению к другому слову из некоторой $T_i \in T^S$.

Множество признаков M^S включает в себя подмножества, обозначаемые далее посредством M с соответствующим нижним индексом и содержащие:

– указания на основу синтаксически главного слова (M_1);

– указания на флексию главного слова (M_2);

– связи “основа – флексия” для синтаксически главного слова (M_3);

– сочетания флексий зависимого и главного слова (M_4). После флексии главного слова через двоеточие при необходимости указывается предлог для связи главного слова с зависимым;

– указания на флексию зависимого слова (M_5).

Посредством $I^S \subseteq G^S \times M^S$ отношения из R разбиваются на классы по сходству основы главного, флексии зависимого слова, лексической и флективной сочетаемости. Для численной оценки схожести СЯУ выполняется редукция ФК (16) исключением объектов и признаков РПЗ согласно следующей теореме.

Теорема 11. Пусть $\{m_1, m_2, m_3\} \subset M_1^S$. Если считать m_1, m_2 и m_3 взаимно различными, то m_1 соответствует указанию на основу главного, m_2 – зависимого слова РПЗ, а m_3 – указанию на основу однословного эквивалента РПЗ при выполнении трех условий:

1. $\exists g_1 \in G^S : I^S(g_1, m_1) = true$, $I^S(g_1, m_3) = false$, $m_2 = p_{bs} \bullet g_1$. Здесь p_{bs} есть обозначение символьной константы *“главное – основа:”*.
2. $\exists \{g_2, g_3\} \subset G^S$, при этом объекты g_1, g_2 и g_3 взаимно различаются, а
$$I^S(g_2, m_3) \wedge I^S(g_3, m_3) \wedge$$
$$\wedge \left(I^S(g_2, m_1) \wedge I^S(g_3, m_2) \vee I^S(g_2, m_2) \wedge I^S(g_3, m_1) \right) = true.$$
3. Не существует других троек объектов, для которых признак m_3 занимал бы место либо m_1, либо m_2 в вышеуказанных соотношениях.

Помимо редукции формальных контекстов отдельных СЯУ, для численной оценки схожести последних вводится аналогичная модель тезауруса ПО:

$$K^{TH} = \left(G^{TH}, M^{TH}, I^{TH} \right), \qquad (17)$$

где множество G^{TH} состоит из символьных пометок отдельных СЯУ. Множество M^{TH} включает элементы множеств признаков формальных контекстов вида (16) всех $g^{TH} \in G^{TH}$. Кроме того, в составе M^{TH} выделяются:

– множество указаний на основы слов, синтаксически подчиненных другим словам в ЕЯ-описаниях ситуаций $g^{TH} \in G^{TH}$. Данное множество обозначается далее как M_6, содержит указания на объекты формальных контекстов вида (16), генерируемых для элементов G^{TH};

– множество связей “основа – флексия” для зависимого слова, M_7;

– множество сочетаний основ зависимого и главного слова, M_8.

Пусть S_1 – ситуация вида (1), соответствующая заведомо корректному ЕЯ-описанию некоторого факта заданной ПО. Положим также, что S_2 – анализируемая СЯУ. Обозначим формальные контексты вида (16): для ситуации S_1 – как K^E, а для ситуации S_2 – как K^X, где $K^E = \left(G^E, M^E, I^E\right)$ и $K^X = \left(G^X, M^X, I^X\right)$, $I^E \subseteq G^E \times M^E$ и $I^X \subseteq G^X \times M^X$, соответственно. Введем обозначения для констант: p_{fl} – для “флексия:”, p_b – для “основа:”. Результат объединения M_6, M_7, M_8, M_4^E, M_4^X, M_5^E и M_5^X, обозначим как M^U.

Определение 6. Будем считать, что ситуации S_1 и S_2 связаны отношением схожести, если каждому объекту $g^X \in G^X$ соответствует такой объект $g^E \in G^E$, что выполняется одно из следующих условий:

(1) $g^X = g^E$ и любой признак $m^E \in M^E$ объекта g^E относится и к g^X.

(2) $g^X = g^E$, при этом *условие (1)* не выполняется, но существует $g^{TH} \in G^{TH}$, обладающий признаком $m_1^{TH} \in M_6$: $m_1^{TH} = p_b \bullet g^E$ при обязательном выполнении следующих условий:

$\left(\exists\, m_{fl}^E \in M_5^E : m_{fl}^E = p_{fl} \bullet f^E\right) \rightarrow \left(\exists\, m_{17}^{TH} \in M_7 : m_{17}^{TH} = g^E \bullet ":" \bullet f^E\right)$,

при этом $\left(I^E\left(g^E, m_{fl}^E\right) \wedge I^X\left(g^E, m_{fl}^E\right)\right) \rightarrow I^{TH}\left(g^{TH}, m_{17}^{TH}\right)$;

$\left(\exists\, m_{bs}^E \in M_1^E : m_{bs}^E = p_{bs} \bullet b^E\right) \rightarrow \left(\exists\, m_{18}^{TH} \in M_8 : m_{18}^{TH} = g^E \bullet ":" \bullet b^E\right)$,

при этом $I^E\left(g^E, m_{bs}^E\right) \rightarrow I^{TH}\left(g^{TH}, m_{18}^{TH}\right)$;

$\left(\exists\, m_{bs}^X \in M_1^X : m_{bs}^X = p_{bs} \bullet b^X\right) \rightarrow \left(\exists\, m_{28}^{TH} \in M_8 : m_{28}^{TH} = g^E \bullet ":" \bullet b^X\right)$,

при этом $I^X\left(g^E, m_{bs}^X\right) \rightarrow I^{TH}\left(g^{TH}, m_{28}^{TH}\right)$.

Кроме того, для $\forall\, m^{TH} \in \left(M^{TH} \setminus M^U\right)$ истинно:

$$I^{TH}\left(g^{TH}, m^{TH}\right) \rightarrow \left(I^E\left(g^E, m^{TH}\right) \wedge I^X\left(g^E, m^{TH}\right)\right). \qquad (18)$$

(3) $g^X \neq g^E$, но существует объект $g^{TH} \in G^{TH}$, обладающий признаками $m_1^{TH} \in M_6$: $m_1^{TH} = p_b \bullet g^E$ и $m_2^{TH} \in M_6$: $m_2^{TH} = p_b \bullet g^X$, при этом для любого признака $m^{TH} \in \left(M^{TH} \setminus M^U\right)$ справедливо:

$$I^{TH}\left(g^{TH}, m^{TH}\right) \rightarrow \left(I^E\left(g^E, m^{TH}\right) \wedge I^X\left(g^X, m^{TH}\right)\right). \qquad (19)$$

(4) $g^X \neq g^E$, но существует объект $g_1^{TH} \in G^{TH}$, обладающий признаком $m_1^{TH} \in M_6$: $m_1^{TH} = p_b \bullet g^E$, а для $\forall\, m^E \in \left(M_4^E \cup M_5^E\right)$ верно:

$$\left(I^{TH}\left(g_1^{TH}, m_1^{TH}\right) \wedge I^E\left(g^E, m^E\right)\right) \rightarrow I^{TH}\left(g_1^{TH}, m^E\right).$$

При этом существуют признаки $m_2^{TH} \in M_6$: $m_2^{TH} = p_b \bullet g^{X_1}$ и $m^X \in \left(M_1^X \cup M_2^X \cup M_3^X\right)$, для которых верно:

$$\left(I^{TH}\left(g_1^{TH}, m_2^{TH}\right) \wedge I^{X}\left(g^{X}, m^{X}\right)\right) \rightarrow I^{TH}\left(g_1^{TH}, m^{X}\right),$$

где $g^{X_1} \neq g^{X}$, а пара $\left(g^{X_1}, g^{E}\right)$ отвечает *условию (3)* при генерации ФК вида (16) для объекта g_1^{TH}. В то же время существует объект $g_2^{TH} \in G^{TH}$, относительно которого пара $\left(g^{X}, g^{X_1}\right)$ также будет отвечать *условию (3)* настоящего *определения*. Генерируемый при этом ФК вида (16) для g_2^{TH} обозначим как K^{X_1}, $K^{X_1} = \left(G^{X_1}, M^{X_1}, I^{X_1}\right)$.

Замечание. Численная оценка схожести СЯУ S_1 и S_2 включает сравнение последовательностей двух и более соподчиненных слов. Выполнимость *определения 6* анализируется только для главных слов. Последовательности считаются заменяемыми, если возможно их построение по формальному контексту (17) на наборе признаков с префиксом p_{bs} для одной и той же СЯУ.

С учётом сопоставления согласно *определению 6* объектов формальных контекстов $K^{E} = \left(G^{E}, M^{E}, I^{E}\right)$ и $K^{X} = \left(G^{X}, M^{X}, I^{X}\right)$, из которых удалена информация РПЗ, *схожесть* ситуаций S_1 и S_2 *численно оценивается* как

$$spc(S_1, S_2) = \frac{\sum_{k=1}^{n} spc_k}{n}, \tag{20}$$

где $n = \left|G^{X}\right|$, а spc_k есть численное значение схожести объектов в паре $\left(g_k^{X}, g^{E}\right)$. В зависимости от выполнимости условий *определения 6* значение spc_k:

– равно 1,0, если для пары $\left(g_k^{X}, g^{E}\right)$ выполнено *условие (1)*;

– вычисляется по формуле:

$$-\log_2\left(1 - \frac{D_c}{path_C}\right) \times \frac{\left|B^{LCS}\right|}{\left|B_1 \setminus B^{LCS}\right| + \left|B_2 \setminus B^{LCS}\right| + \left|B^{LCS}\right|}, \tag{21}$$

если для пары $\left(g_k^{X}, g^{E}\right)$ выполнено *условие (2), (3)* либо *(4)*.

Во втором случае имеем гипотетическую решетку ФП (обозначим ее как $\Re^{XE}$), в которой объемы объектных ФП (ФП с одним объектом в составе объема) есть $\left\{g_k^{X}\right\}$ и $\left\{g^{E}\right\}$ (при выполнении *условия (2)* или *(3)*) либо $\left\{g_k^{X}\right\}$, $\left\{g^{E}\right\}$ и $\left\{g^{X_1}\right\}$ (при выполнении *условия (4)*). Значение D_c равно числу сравнимых ФП, составляющих цепочку с вершинным ФП решетки $\Re^{XE}$ в качестве максимального ФП и Наименьшим Общим Суперпонятием (НОСП) для объектных ФП решетки $\Re^{XE}$ – в качестве минимального ФП. Множество B^{LCS} есть содержание (множе-

ство признаков всех объектов) этого НОСП, а число $path_C$ равно минимальному числу ФП в цепочке, которой принадлежит вершинное ФП, наименьшее ФП решетки $\mathfrak{R}^{XE}$ и формальное понятие с содержанием B^{LCS}.

В случае выполнения любого из *условий (2), (3)* или *(4)* значение $D_c = 2$.

При выполнении *условия (2)* либо *(3)* $path_C = 4$, а в B^{LCS} войдут признаки $m^{TH} \in \left(M^{TH} \setminus M^{U}\right)$, для каждого из которых справедливо либо соотношение (18) (при выполнении *условия (2)*), либо соотношение (19) (при выполнении *условия (3)*). Множества B_1 и B_2 в этом случае определяются следующим образом:

$$B_1 = \left\{ m^E : m^E \in \left(M_1^E \cup M_2^E \cup M_3^E\right),\ I^E\left(g^E, m^E\right) = true \right\},$$
$$B_2 = \left\{ m^X : m^X \in \left(M_1^X \cup M_2^X \cup M_3^X\right),\ I^X\left(g_k^X, m^X\right) = true \right\}.$$

Доказательство выполнимости *условия (4)* обычно происходит в несколько итераций. При этом в ходе каждой последующей итерации число признаков, не являющихся общими для g_k^X и g^{X_1}, всегда меньше, чем в предыдущей. Начальное значение $path_C$, равное 4, в ходе каждой итерации увеличивается на 1, а

$$B_1 = \left\{ m^{X_1} : m^{X_1} \in \left(M_1^{X_1} \cup M_2^{X_1} \cup M_3^{X_1}\right),\ I^{X_1}\left(g^{X_1}, m^{X_1}\right) = true \right\},$$
$$B_2 = \left\{ m^X : m^X \in \left(M_1^{X_1} \cup M_2^{X_1} \cup M_3^{X_1}\right),\ I^{X_1}\left(g_k^X, m^X\right) = true \right\},$$

где $\left(M_1^{X_1} \cup M_2^{X_1} \cup M_3^{X_1}\right) \subset M^{X_1}$, а $B^{LCS} = B_1 \cap B_2$.

Далее в главе приводится пример интерпретации теста открытой формы с вычислением оценок (20) для каждого ответа. Пусть СЯУ S_1 задана четырьмя предложениями, представляющими правильный ответ на вопрос о связи переобучения и эмпирического риска. Допустим, имеются три варианта СЯУ S_2 (см. табл. 1), связанные отношением схожести с S_1 согласно *определению 6*.

Таблица 1

Сравнение ответов с эталоном

ответы	эталон				анализируемый		
вариант	1	2	3	4	1	2	3
основа	флективная часть + предлог						
заниженн	ости	ости	ость	ость	ость	ость	ости
эмпирическ	ого	ого	ого	ого	–	–	–
риск	а	а	а	а	–	–	–
средн	–	–	–	–	ей	ей	ей
ошибк	–	–	–	–	и:на	и:на	и:на
обучающ	–	–	–	–	ей	ей	ей
выборк	–	–	–	–	е	е	е
переобучении	е	–	–	ем	ем	–	е
переподгонк	–	а	ой	–	–	ой	–
связан	–	–	а:с	а:с	а:с	а:с	–
привод	ит:к	ит:к	–	–	–	–	ит:к

Фрагмент тезауруса ПО “Математические методы обучения по прецедентам”, задействованный в доказательстве и численной оценке схожести СЯУ, представлен в табл. 2 ЕЯ-описанием соответствующих фактов.

Таблица 2

Факты ПО для фрагмента тезауруса

№ п/п	1		2		3		4	
Основа	Флективная часть + предлог							
заниженн	ость	ость	ости	ости	–	ость	ости	ость	ость
оценк	–	–	–	–	–	и	и	и	и
эмпирическ	ого	–	ого	–	–	–	–	–	–
риск	а	–	а	–	–	–	–	–	–
средн	–	ей	–	ей	–	–	–	–	–
ошибк	–	и:на	–	и:на	–	–	–	и	и
распознавани	–	–	–	–	–	–	–	я	я
обучающ	–	ей	–	ей	–	–	–	–	–
выборк	–	е	–	е	–	–	–	–	–
переусложнени	ем	ем	е	е	–	–	–	–	–
модел	и	и	и	и	–	–	–	–	–
уменьшени	–	–	–	–	е	–	–	–	–
обобщающ	–	–	–	–	ей	ей	ей	–	–
способност	–	–	–	–	и	и	и	–	–
выбор	–	–	–	–	–	–	–	ом	а
решающ	–	–	–	–	его	–	–	его	его
дерев	–	–	–	–	а	–	–	–	–
правил	–	–	–	–	–	–	–	а	а
алгоритм	–	–	–	–	–	а	а	–	–
переподгонк	–	–	–	–	ой	ой	а	–	–
переобучени	–	–	–	–	–	ем	е	–	–
связан	а:с	а:с	–	–	о:с	а:с	–	а:с	–
вызван	а	а	–	–	–	а	–	–	–
обусловлен	а	а	–	–	о	–	–	–	–
привод	–	–	ит:к	ит:к	–	–	ит:к	–	–
завис	–	–	–	–	–	–	–	–	ит:от

Использованные в эксперименте формальные контексты строились по результатам синтаксического разбора ЕЯ-фраз программой “Cognitive Dwarf”.

Таблица 3

Численная оценка близости ответа эталону

Вариант	$spc(S_1, S_2)$	$\left\|B^{LCS}\right\|$	$\left\|B_1 \setminus B^{LCS}\right\|$	$\left\|B_2 \setminus B^{LCS}\right\|$
1	0,9167	7,7500	0,7500	0,0000
2	0,7917	7,0000	2,0000	0,5000
3	0,8750	7,7500	0,7500	0,7500

Основные результаты главы представлены в работах [1, 17, 26, 42, 43] из списка публикаций автора.

Шестая глава диссертации посвящена оптимальной организации ТЗОФ на основе предложенных в предыдущей главе методов и моделей. Вводится понятие смыслового эталона СЯУ и рассматриваются два приближенных метода его построения с представлением формальным контекстом вида (16).

Первый метод основан на подходе к выделению и классификации синтагматических зависимостей, предложенном в **разделе 3.5**.

Пусть $K^E = \left(G^E, M^E, I^E\right)$ есть искомый формальный контекст эталона. Если $\{j,k\} \subset J$ и $\exists E \in V^J : (j,k) \in E$ в дереве (13), расширенном с учетом условий *леммы 3* и *теоремы 8*, то для основ b_j и b_k и флексий f_j и f_k соответствующие им элементы множеств G^E и M^E, а также элементы отношения I^E, будут сформированы следующим образом.

Случай 1. Индекс k соответствует родительскому узлу, индекс j – дочернему узлу в расширенном дереве (13), а линейная структура ЕЯ-фразы не содержит предлог между словами с индексами j и k.

При этом в состав множества признаков M^E ФК $K^E = \left(G^E, M^E, I^E\right)$ будут включены признаки $m_1 = p_{bs} \bullet b_k$, $m_2 = p_{bf} \bullet f_k$, $m_3 = p_{fl} \bullet f_j$ и $m_4 = f_j \bullet ":" \bullet f_k$, основа b_j войдет в множество объектов G^E указанного ФК, а пары (b_j, m_1), (b_j, m_2), (b_j, m_3) и (b_j, m_4) войдут в отношение I^E.

Случай 2. Индекс k соответствует родительскому узлу, индекс j – дочернему узлу в расширенном дереве (13), линейная структура ЕЯ-фразы содержит предлог p_y между словами с индексами j и k.

В этом случае признаки m_1 и m_3 формируются аналогично *случаю 1*, $m_2 = p_{bf} \bullet f_k \bullet ":" \bullet p_y$, $m_4 = f_j \bullet ":" \bullet f_k \bullet ":" \bullet p_y$, пары (b_j, m_1), (b_j, m_2), (b_j, m_3) и (b_j, m_4) включаются в отношение I^E.

Второй метод основан на построении ФК эталона по совокупности формальных контекстов вида (16) для отдельных СЭ-фраз, задающих СЯУ. При этом формальные контексты указанной совокупности строятся по результатам разбора этих фраз внешней программой синтаксического анализа. Для отбора объектов и признаков из формальных контекстов отдельных фраз вводятся коэффициенты сжатия информации относительно формального контекста вида (16).

Коэффициент сжатия информации по основам равен:

$$k^S = \frac{\sum_{i=1}^{n^{BS}} k_i^S}{n^{BS}}, \qquad (22)$$

где $k_i^S = \frac{\sum_{j=1}^{n_i^{BS}} \sum_{k=1}^{n^{MF}} n_{ijk}^{AS}}{n_i^{BS}}$; $n^{BS} = |M_1|$; $n^{MF} = |M_2|$;

$$n_i^{BS} = \left| \left\{ g \in G^S : I^S(g,m) = true,\ m \in M_1,\ m = p_{bs} \bullet b_i \right\} \right|;$$

$$n_{ijk}^{AS} = \left| \left\{ m_k \in M_3 : I^S(g,m_k) = true, \exists\, m_{bf} \in M_2,\ m_{bf} = p_{bf} \bullet f_k,\ m_k = b_i \bullet ":" \bullet f_k \right\} \right|;$$

p_{bf} соответствует символьной константе *“главное – флексия:”*.

Аналогично определяется коэффициент сжатия информации по флексиям:

$$k^F = \frac{\sum_{i=1}^{n^{FS}} k_i^F}{n^{FS}}, \qquad (23)$$

где $k_i^F = \dfrac{\sum_{j=1}^{n_i^{FS}} \sum_{k=1}^{n^{MF}} n_{ijk}^{AF}}{n_i^{FS}}$; $n^{FS} = |M_5|$;

$$n_i^{FS} = \left| \left\{ g \in G^S : I^S(g,m) = true,\ m \in M_5,\ m = p_{fl} \bullet f_i \right\} \right|;$$

$$n_{ijk}^{AF} = \left| \left\{ m \in M_4 : I^S(g_j,m) = true,\ \exists\, m_{bf} \in M_2,\ m_{bf} = p_{bf} \bullet f_k,\ m = f_i \bullet ":" \bullet f_k \right\} \right|.$$

В главе представлен алгоритм построения ФК эталона, реализующий отбор объектов и признаков из формальных контекстов отдельных фраз по максимуму коэффициентов (22) и (23) результирующего ФК. При этом признак будет включен в множество признаков формального контекста эталона, если он входит в состав пятерки признаков $\{m_1, m_2, m_3, m_4, m_5\}$, в которой $m_1 = p_{bs} \bullet b$, $m_2 = p_{bf} \bullet f_1$, $m_3 = b \bullet ":" \bullet f_1$, $m_4 = p_{fl} \bullet f_2$, $m_5 = f_2 \bullet ":" \bullet f_1$. При этом основе b не должен соответствовать объект ФК, если есть другой объект этого же ФК, который обладает одновременно признаком m_1 и некоторым другим признаком $m = p_{bs} \bullet b_1$, где $b_1 \neq b$, а основе b_1 не соответствует ни одного объекта этого ФК при том, что признак m относится более чем к одному объекту.

Замечание. Последовательности из трех и более соподчиненных слов, встречающиеся более чем в 49% исходных СЭ-фраз, выделяются предварительно на этапе синтаксического разбора. Для каждой такой последовательности строится свой ФК вида (16), который будет объединен с формальным контекстом эталона. Данный шаг предпринят в целях нежелательного занижения коэффициентов (22) и (23) при выполнении рассматриваемого алгоритма.

Вне зависимости от способа формирования смыслового эталона *точность решения численно оценивается* средним числом невыделенных (опущенных) признаков на один объект формального контекста сформированного эталона. Значение данного показателя будет тем выше, чем меньше частота, с которой сочетания слов в основе отношения “объект-признак” для ФК эталона совместно встречаются в различных СЭ-фразах.

Качественно процесс формирования смысловых эталонов в целом характеризуется соотношением размеров тезауруса, задаваемого моделью (17), при построении его на основе формальных контекстов вида (16) для всех СЭ-фраз ка-

ждой СЯУ и на основе эталонов при заданном числе СЯУ в тезаурусе. Пример указанного соотношения приведен на рис. 1 для СЯУ из табл. 4. Часть указанных СЯУ была задействована при построении тезауруса в табл. 2.

Таблица 4

Ситуации языкового употребления

i	Что описывает СЯУ
1	Связь переобучения с эмпирическим риском
2	Связь переусложнения модели с заниженностью средней ошибки на тренировочной выборке
3	Влияние переподгонки на частоту ошибок дерева принятия решений
4	Причина заниженности оценки обобщающей способности алгоритма
5	Зависимость оценки ошибки распознавания от выбора решающего правила
6	Зависимость обобщающей способности логического алгоритма классификации от числа закономерностей алгоритмической композиции

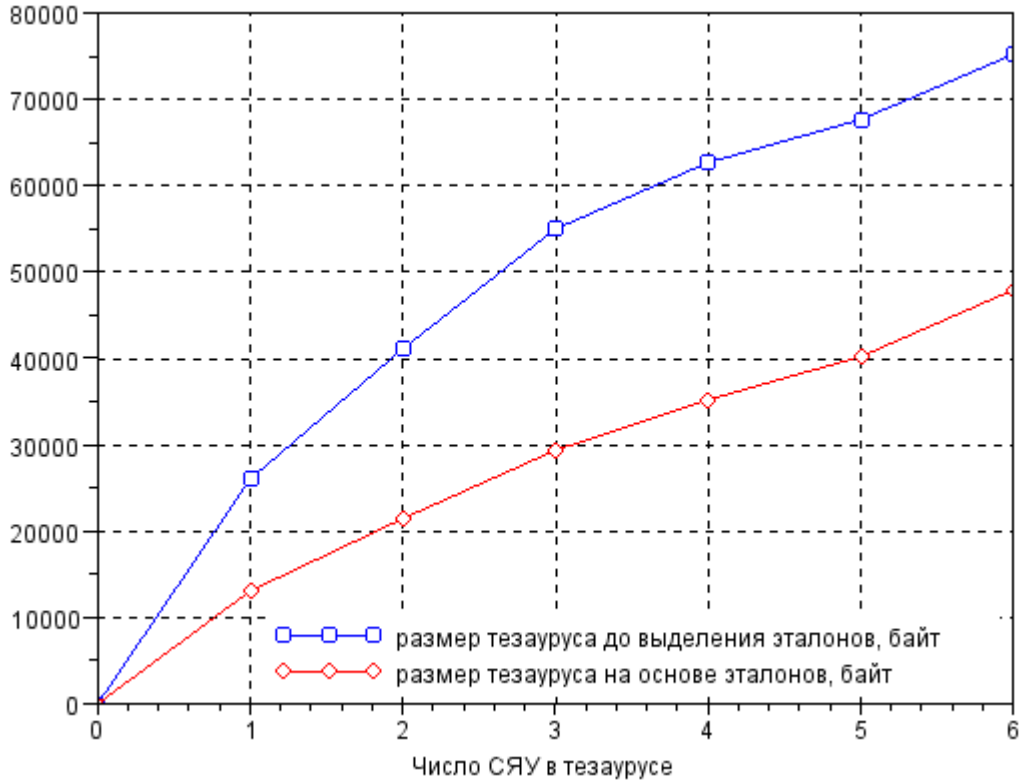

Рис. 1. Размер тезауруса для разного числа СЯУ

Для сравнения в табл. 5 приведены значения числа СЭ-фраз, задающих отдельную СЯУ (N_1),числа фраз, представляющих эталон (N_2), исходного числа

объектов (N_3) и признаков СЯУ (N_4), числа объектов (N_5) и признаков эталона (N_6).

Таблица 5

Смысловые эталоны

i	1	2	3	4	5	6
$N_1(i)$	54	53	26	26	2	3
$N_2(i)$	14	15	5	11	2	3
$N_3(i)$	13	15	13	12	8	11
$N_4(i)$	160	153	135	102	46	68
$N_5(i)$	9	12	12	12	8	11
$N_6(i)$	75	78	65	71	46	68

В сформированном ФК эталона все обозначения основ слов в составе имен объектов и признаков заменяются переменными, а для каждой переменной задается конкретизация некоторой основой. Преобразованный указанным образом ФК есть шаблон формального контекста эталона. Аналогичные замены производятся для каждой ЕЯ-фразы исходного множества, где отдельное слово при этом представлено парой “основа-флексия” по результатам выделения эталона. Пусть T^P есть множество последовательностей указанных пар, в которых основы заменены переменными. Тогда интерпретация ответа на ТЗОФ в значительном числе случаев есть “наложение” ответа на один из элементов множества T^P с формированием списков конкретизаций переменных и последующим сравнением с аналогичными списками для “правильного” ответа. Сама интерпретация происходит за линейное время, пропорциональное $\left|T^P\right|$. Далее в главе показано, каким образом совокупность шаблонов формальных контекстов известных смысловых эталонов может быть использована для построения потенциально возможных эталонов на совокупности СЯУ по заданной ПО.

Завершает главу описание архитектуры системы контроля знаний с применением ТЗОФ. На рис. 2 представлены основные компоненты системы:

- БФЭ – блок формирования эталонов;
- БФШ – блок формирования шаблонов;
- БСШ – блок слияния шаблонов;
- БФТ – блок формирования тезауруса;
- БВТ – блок выбора теста;
- ТЕСТ – блок выполнения теста;
- БФЗ – блок формирования заданий, помещаемых в базу ЗАДАНИЯ;

– компоненты ТЕЗАУРУС, КОНКРЕТИЗАЦИИ и ШАБЛОНЫ составляют базу предметно-языковых знаний системы. Сюда же входит СО – база Синтаксических Отношений, формируемых в БФО – блоке формирования отношений на основе базы шаблонов формальных контекстов эталонов.

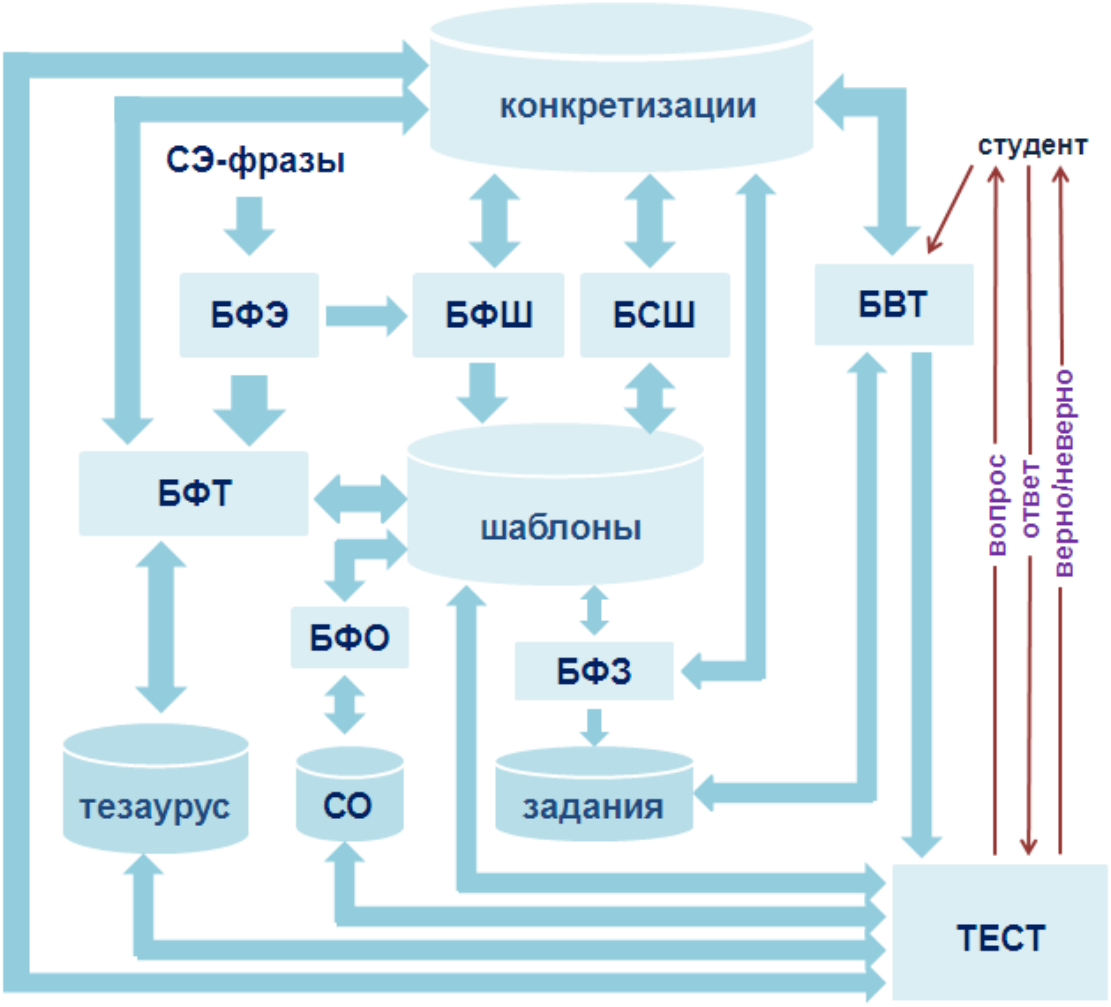

Рис. 2. Архитектура программной системы тестирования знаний

Каждое ТЗОФ представляет собой совокупность вопроса и шаблона ФК эталона для “правильного” ответа плюс соответствующие конкретизации для переменных, представляемые базой КОНКРЕТИЗАЦИИ. В соответствии с выбранными заданиями блок ТЕСТ реализует саму процедуру тестирования. После ввода текста ответа система делает попытку применить шаблон “правильного” ответа с учетом конкретизаций, заданных для шаблона в рамках задания. Если сопоставление неуспешно, делается попытка применить другие шаблоны из базы и в случае успеха – доказать схожесть между СЯУ для ответа испытуемого и для “правильного” ответа. При успешном доказательстве вычисляется оценка схожести по формуле (20), а полученное значение может быть использовано при выставлении испытуемым оценок, а также для сбора статистики.

Основные результаты главы отражены в работах [8, 17, 23, 25, 44, 51, 55, 60] из списка публикаций автора.

Заключение

Основные научные результаты в области *разработки новых математических методов моделирования объектов и явлений* состоят в следующем.

1. Разработана *теоретико-множественная модель* процесса установления семантической эквивалентности для флективного языка в рамках синонимического варьирования на уровне абстрактной лексики.

Введение в рассмотрение конечного множества корректно формализуемых правил синонимических преобразований на основе аппарата стандартных лексических функций даёт возможность численно оценивать схожесть смыслов высказываний естественного языка не зависящим от их предметной области способом и с учетом большинства возможных случаев синонимии.

2. Предложена *математическая модель* и реализующий её *комплекс программ* формирования и кластеризации смысловых отношений на основе описаний ситуаций действительности множествами эквивалентных по смыслу фраз предметно-ограниченного подмножества естественного языка.

Новизна решения заключается в сравнении символьных последовательностей, составляющих *эквивалентные по смыслу* описания одного и того же *объекта* (ситуации) на заданном языке, с выделением изменяемых и неизменяемых частей для последующего анализа взаимного расположения фрагментов последовательностей в языковых конструкциях с разными логическими акцентами относительно одной и той же ситуации. Предложенная модель выявления закономерностей сосуществования словоформ в линейном ряду выделяет для заданного естественного языка лучший способ выражения нужной мысли, который составляет основу смыслового эталона. Сказанное актуально как для разработки стратегий и правил синтаксического анализа, так и для ролевой идентификации сущностей при формировании признаков сравниваемых текстов.

3. Предложен *метод численной оценки* смысловой *близости текстов* предметного языка для интерпретации результатов теста открытой формы.

Новизной метода является *теоретико-решеточная модель* ситуации языкового употребления в качестве информационной единицы тезауруса. При этом *смысловая близость* ответа испытуемого эталонному ответу *оценивается числом* признаков, которые характеризуют сочетаемость слов и разделяются *объектами* сравниваемых СЯУ относительно тезауруса.

4. Предложены два приближённых альтернативных *метода* автоматизированного *построения модели смыслового эталона* в виде *решётки формальных понятий* (включая численные оценки точности получаемого решения), *модель* процесса интерпретации ответа испытуемого на тестовое задание открытой формы, а также *метод компрессии текстовой базы знаний* на основе выделенных эталонов.

Вне зависимости от пути формирования эталона его выделение сокращает размер базы знаний для численной оценки близости текста эталону в среднем на 40–50%. Предложенные методы построения эталона и комплекс программ формирования смысловых отношений закладывают основу *нового научного направления* – сжатия информации без потери смысловой составляющей.

5. Предложена типовая *архитектура программной системы* контроля знаний, реализующая предложенные в работе методы и модели.

В области *развития качественных и приближённых аналитических методов исследования математических моделей* основной научный результат есть предложенное в работе решение задачи построения системы целевых выводов в грамматике деревьев (Δ-грамматике).

В отличие от традиционных подходов к формализации преобразований помеченных деревьев, с целью нахождения последовательности преобразований с заданными свойствами было предложено *исследовать динамику функционирования совокупности правил Δ-грамматики на основе ее информационно-логической модели*. Разработанная модель учитывает недетерминированный характер порождения множества помеченных деревьев, а построение целевого вывода сводится к классическим задачам сетей Петри.

Важнейшие результаты в области *комплексных исследований научных и технических проблем с применением современной технологии математического моделирования и вычислительного эксперимента* состоят в следующем.

1. На основе теории *анализа формальных понятий* предложен *комплексный подход* к решению задачи пополнения лингвистических информационных ресурсов из текстов с последующим упорядочиванием знаний.

За счёт использования формального понятия в качестве базового элемента информационного ресурса предложенная модель тезауруса в виде решетки формальных понятий оперирует данными на семантическом уровне без потери или недопустимого упрощения объектов и их признаков.

2. Разработан *комплексный подход* к решению задачи *формирования и кластеризации понятий* на основе синтаксического контекста существительного русского языка.

Синтаксический разбор на основе наиболее вероятных связей в совокупности с методами АФП даёт высокую (менее 2% ошибок) точность выделения связей “объект – признак” независимо от ограничений на перифразирование.

В приложении приведены фрагменты текста на Visual Prolog 5.2 программы формирования модели СЯУ, выделяющей классы смысловых отношений в основе эталона и синтаксические отношения как их частный случай. Пример на рис.1, табл.4 и 5 иллюстрируют достигаемое сокращение размеров базы знаний при выделении смысловых эталонов реализованными в работе методами.

Список основных публикаций автора по теме диссертации

Монография

1. Михайлов Д.В. Теоретические основы построения открытых вопросно-ответных систем. Семантическая эквивалентность текстов и модели их распознавания: монография / Д.В. Михайлов, Г.М. Емельянов. Великий Новгород: НовГУ им. Ярослава Мудрого, 2010. 286 с. 16,1 п.л. (вклад автора – 14,5 п.л.)

Статьи в рецензируемых научных журналах, включенных в реестр ВАК МОиН РФ

2. Емельянов Г.М. Распознавание сверхфразовых единств при установлении эквивалентности смысловых образов высказываний в общей задаче моделирования языковой деятельности / Г.М. Емельянов, Д.В. Михайлов // Известия СПбГЭТУ “ЛЭТИ”, сер. “Информатика, управление и компьютерные технологии”. СПб., 2003. Вып. 1. С. 65–73. 0,52 п.л. (вклад автора – 0,47 п.л.)

3. Михайлов Д.В. Информационно-логическая модель системы правил Δ-грамматики / Д.В. Михайлов, Г.М. Емельянов // Известия СПбГЭТУ “ЛЭТИ”, сер. “Информатика, управление и компьютерные технологии”. СПб., 2003. Вып. 3. С. 96–102. 0,70 п.л. (вклад автора – 0,63 п.л.)

4. Михайлов Д.В. Построение модели объекта информационного пространства применительно к исследованию динамики функционирования Δ-грамматик / Д.В. Михайлов, Г.М. Емельянов // Вестник Новгородского государственного университета имени Ярослава Мудрого, сер. “Технические науки”. 2004. № 26. С. 131–136. 0,70 п.л. (вклад автора – 0,63 п.л.)

5. Михайлов Д.В. Представление смысла в задаче установления семантической эквивалентности высказываний / Д.В. Михайлов, Г.М. Емельянов // Вестн. Новгородского гос. ун-та им. Ярослава Мудрого, сер. “Технические науки”. 2004. № 28. С. 106–110. 0,58 п.л. (вклад автора – 0,52 п.л.)

6. Корнышов А.Н. Концептуально-ситуационное моделирование высказываний естественного языка в задаче анализа их смысловой эквивалентности / А. Н. Корнышов, Д.В. Михайлов // Вестн. Новгородского гос. ун-та им. Ярослава Мудрого, сер. “Технические науки”. 2005. № 34. С. 76–80. 0,58 п.л. (вклад автора – 0,15 п.л.)

7. Михайлов Д.В. Семантическая кластеризация текстов предметных языков (морфология и синтаксис) / Д.В. Михайлов, Г.М. Емельянов // Компьютерная оптика. 2009. Т. 33, № 4. С. 473–480. 0,26 п.л. (вклад автора – 0,23 п.л.)

8. Emelyanov G.M. Development of Recognition System of Analysis of Semantic Images of Natural Language Statements / G.M. Emelyanov, E.I. Zaitseva, D.V. Mikhailov, E.P. Kurashova // Pattern Recognition and Image Analysis. 2003. Vol. 13. N 2. P. 251–253. 0,37 п.л. (вклад автора – 0,22 п.л.)

9. Emelyanov G. M. Synonymic Transformations in Analysis of Semantic Pattern Equivalence at the Superphrase Unity Level / G. M. Emelyanov, D. V. Mikhailov, E. I. Zaitseva // Pattern Recognition and Image Analysis. 2003. Vol. 13. N 1. P. 21–23. 0,37 п.л. (вклад автора – 0,24 п.л.)

10. Emelyanov G. M. Recognition of Superphrase Unities in Texts while Establishing Their Semantic Equivalence / G. M. Emelyanov, D. V. Mikhailov, E. I. Zaitseva // Pattern Recognition and Image Analysis. 2003. Vol. 13. N 3. P. 447–451. 0,61 п.л. (вклад автора – 0,52 п.л.)

11. Emelyanov G. M. Semantic Relation Analysis for Classification of the Meaning Patterns of Utterances / G. M. Emelyanov, D. V. Mikhailov, N. A. Stepanova // Pattern Recognition and Image Analysis. 2005. Vol. 15. N 2. P. 382–383. 0,24 п.л. (вклад автора – 0,06 п.л.)

12. Emelyanov G. M. Updating the Language Knowledge Base in the Problem of Equivalence Analysis of Semantic Images of Statements / G. M. Emelyanov, D. V. Mikhailov // Pattern Recognition and Image Analysis. 2005. Vol. 15. N 2. P. 384–386. 0,37 п.л. (вклад автора – 0,33 п.л.)

13. Emel'yanov G. M. Filling in the Government-Pattern Dictionary in the Analysis of Equivalence for Sense Images of Statements / G. M. Emel'yanov, D. V. Mikhailov // Pattern Recognition and Image Analysis. 2007. Vol. 17. N 2. P. 268–273. 0,73 п.л. (вклад автора – 0,66 п.л.)

14. Emel'yanov G. M. Analysis of Semantic Relations in Classification of Sense Images of Statements / G. M. Emel'yanov, D. V. Mikhailov, N. A. Stepanova // Pattern Recognition and Image Analysis. 2007. Vol. 17. N 2. P. 274–278. 0,61 п.л. (вклад автора – 0,20 п.л.)

15. Emel'yanov G. M. Clusterization of Semantic Meanings in the Problem of Sense Equivalence Situation Recognition / G. M. Emel'yanov, D. V. Mikhailov // Pattern Recognition and Image Analysis. 2009. Vol. 19. N 1. P. 92–102. 1,34 п.л. (вклад автора – 1,21 п.л.)

16. Mikhailov D. V. Formation and clustering of noun contexts within the framework of Splintered Values / D. V. Mikhailov, G. M. Emelyanov, N. A. Stepanova // Pattern Recognition and Image Analysis. 2009. Vol. 19, N 4. P. 664–672. 1,10 п.л. (вклад автора – 0,88 п.л.)

17. Mikhailov D. V. Semantic Clustering and Affinity Measure of Subject-Oriented Language Texts / D.V. Mikhailov, G.M. Emel'yanov // Pattern Recognition and Image Analysis. 2010. Vol. 20. N 3. P. 376–385. 1,22 п.л. (вклад автора – 1,14 п.л.)

Доклады на международных конференциях

18. Емельянов Г. М. Вопросы моделирования семантической связанности для систем понимания текста / Г. М. Емельянов, Д. В. Михайлов // Распознавание-2001: сб. мат-лов 5-й Междунар. конф. Курск: Курский гуманитарно-техн. инст-т; Курский гос. техн. ун-т, 2001. Ч. 1. С. 56–58. 0,17 п.л. (вклад автора – 0,16 п.л.)

19. Емельянов Г. М. Динамическая модель естественного языка в системах пользовательских интерфейсов / Г. М. Емельянов, Д. В. Михайлов, Е. И. Зайцева // Междунар. конф. по компьютерной лингвистике "Диалог-2002". М.: Наука, 2002. Т. 2. С. 165–170. 0,75 п.л. (вклад автора – 0,56 п.л.)

20. Емельянов Г. М. Динамическая модель естественного языка в системах пользовательских интерфейсов / Г. М. Емельянов, Е. И. Зайцева, Д. В. Михайлов // Интеллектуализация обработки информации: тез. докл. Междунар. науч. конф. Симферополь: Крымский научный центр НАН Украины, Таврический национальный университет, 2002. С. 120–121. 0,06 п.л. (вклад автора – 0,04 п.л.)

21. Емельянов Г. М. Применение аппарата ограниченных сетей Петри для построения динамической модели естественного языка / Г. М. Емельянов, Е. И. Зайцева, Д. В. Михайлов // Интеллектуализация обработки информации:

тез. докл. Междунар. науч. конф. Симферополь: Крымский научный центр НАН Украины, Таврический национальный университет, 2002. С. 121–122. 0,06 п.л. (вклад автора – 0,04 п.л.)

22. Емельянов Г. М. Синонимические преобразования в задаче анализа эквивалентности смысловых образов высказываний на уровне сверхфразовых единств / Г. М. Емельянов, Д. В. Михайлов, Е. И. Зайцева // Распознавание образов и анализ изображений: новые информационные технологии (РОАИ-6-2002): труды 6-й Междунар. конф.; НовГУ им. Ярослава Мудрого. Великий Новгород, 2002. Т. 1. С. 215–219. 0,58 п.л. (вклад автора – 0,38 п.л.)

23. Емельянов Г. М. К разработке распознающей системы анализа смысловых образов высказываний на естественном языке / Г. М. Емельянов, Е. И. Зайцева, Д. В. Михайлов, Е. П. Курашова // Распознавание образов и анализ изображений: новые информационные технологии (РОАИ-6-2002): труды 6-й Междунар. конф.; НовГУ им. Ярослава Мудрого. Великий Новгород, 2002. Т. 1. С. 220–223. 0,47 п.л. (вклад автора – 0,28 п.л.)

24. Михайлов Д. В. Информационное наполнение дерева в задаче исследования динамики функционирования Δ-грамматики / Д. В. Михайлов, Г. М. Емельянов // Распознавание-2003: сб. мат-лов 6-й Междунар. конф. Курск: Курский гос. техн. ун-т, 2003. Ч. 1. С. 35–37. 0,12 п.л. (вклад автора – 0,11 п.л.)

25. Емельянов Г. М. Установление смысловой эквивалентности высказываний: на пути к решению проблемы / Г. М. Емельянов, Д. В. Михайлов // Интеллектуализация обработки информации: тез. докл. Междунар. науч. конф. Симферополь: Крымский научный центр НАН Украины, 2004. С. 70. 0,06 п.л. (вклад автора – 0,05 п.л.)

26. Emelyanov G. M. Application of the computer thesaurus for automation of updating of the Government Patterns's dictionary / G. M. Emelyanov, D. V. Mikhailov, N. A. Stepanova // VI International Congress on Mathematical Modeling. Book of Abstracts; University of Nizhny Novgorod. Nizhny Novgorod, 2004. P. 352. 0,12 п.л. (вклад автора – 0,02 п.л.)

27. Emelyanov G. M. Updating of the language knowledge base in the problem of statement's semantic images's equivalence's analysis / G. M. Emelyanov, D. V. Mikhailov // 7th Int. Conf. on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-7-2004). Conf. Proc. St. Petersburg: SPbETU, 2004. Vol. II. P. 462–465. 0,47 п.л. (вклад автора – 0,42 п.л.)

28. Emelyanov G.M. Semantic relation analysis for classification of meaning pattern of utterances / G.M. Emelyanov, D.V. Mikhailov, N.A. Stepanova // 7th Int. Conf. on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-7-2004). Conf. Proc. St. Petersburg: SPbETU, 2004. Vol. II. P. 460–461. 0,23 п.л. (вклад автора – 0,06 п.л.)

29. Корнышов А. Н. Концептуальный уровень и его использование в задаче моделирования синонимических преобразований высказываний естественного языка / А. Н. Корнышов, Д. В. Михайлов // Математика в вузе: мат-лы XVIII Междунар. науч.-метод. конф. СПб.: Петербургский гос. ун-т путей сообщения, 2005. С. 118–120. 0,17 п.л. (вклад автора – 0,04 п.л.)

30. Mikhailov D. V. Application Of The Predicate Word's Lexical Meanings's System For Automation Of Updating Of The Dictionary Of Government Patterns / D. V. Mikhailov, G. M. Emelyanov // Interactive Systems And Technologies: The Problems of Human-Computer Interaction. Collections of scientific papers. Ulyanovsk: ULSTU, 2005. P. 164–168. 0,40 п.л. (вклад автора – 0,37 п.л.)

31. Михайлов Д. В. Иерархия семантических отношений в задаче построения Модели Управления предикатного слова / Д. В. Михайлов, Г. М. Емельянов // Распознавание-2005: сб. мат-лов 7-й Междунар. конф. Курск: Курский гос. техн. ун-т, 2005. С. 42–43. 0,12 п.л. (вклад автора – 0,11 п.л.)

32. Емельянов Г. М. Концептуально-ситуационное моделирование процесса перифразирования высказываний Естественного Языка как обучение на основе прецедентов / Г. М. Емельянов, А. Н. Корнышов, Д. В. Михайлов // Интеллектуализация обработки информации: тез. докл. Междунар. науч. конф. Симферополь: Крымский научный центр НАН Украины, 2006. С. 78–79. 0,12 п.л. (вклад автора – 0,02 п.л.)

33. Михайлов Д. В. Модель сортовой системы языка в задаче построения семантического образа высказывания на уровне глубинного синтаксиса / Д. В. Михайлов, Г. М. Емельянов // Интеллектуализация обработки информации: тез. докл. Междунар. науч. конф. Симферополь: Крымский научный центр НАН Украины, 2006. С. 148–150. 0,12 п.л. (вклад автора – 0,11 п.л.)

34. Mikhailov D. V. Roles's contents of word's lexical meaning's in a problem of recognition of synonymy's situations on the basis of standard lexical functions / D. V. Mikhailov, G. M. Emelyanov // Interactive Systems And Technologies: The Problems of Human-Computer Interaction. Collections of scientific papers. Ulyanovsk: ULSTU, 2007. P. 159–165. 0,56 п.л. (вклад автора – 0,51 п.л.)

35. Emelyanov G. M. Formalization of the word's lexical meaning in a problem of recognition of natural language's statements's synonymy's situations / G. M. Emelyanov, D. V. Mikhailov // 8th Int. Conf. "Pattern Recognition and Image Analysis: New Information Technologies" (PRIA-8-2007). Conf. Proc. Yoshkar-Ola: Mari State Technical University, 2007. Vol. 2. P. 253–257. 0,58 п.л. (вклад автора – 0,52 п.л.)

36. Корнышов А. Н. Таксономия знаний в задаче распознавания семантических отношений / А. Н. Корнышов, Д. В. Михайлов // Распознавание-2008: сб. мат-лов VIII Междунар. конф. Курск: Курский гос. техн. ун-т, 2008. Ч. 1. С. 183–185. 0,12 п.л. (вклад автора – 0,04 п.л.)

37. Михайлов Д. В. Формирование и кластеризация понятий в задаче автоматизированного построения тезауруса предметной области / Д. В. Михайлов, Г. М. Емельянов // Распознавание-2008: сб. мат-лов VIII Междунар. конф. Курск: Курский гос. техн. ун-т, 2008. Ч. 2. С. 20–22. 0,12 п.л. (вклад автора – 0,11 п.л.)

38. Корнышов А. Н. Иерархизация системы предикатов семантических отношений / А. Н. Корнышов, Д. В. Михайлов // Интеллектуализация обработки информации: тез. докл. Междунар. науч. конф. Симферополь: Крымский научный центр НАН Украины, 2008. С. 130–131. 0,12 п.л. (вклад автора – 0,04 п.л.)

39. Михайлов Д. В. Формирование и кластеризация понятий на основе множества ситуационных контекстов / Д. В. Михайлов, Г. М. Емельянов, Н. А. Степанова // Интеллектуализация обработки информации: тез. докл. Междунар. науч. конф. Симферополь: Крымский научный центр НАН Украины, 2008. С. 168–170. 0,17 п.л. (вклад автора – 0,16 уч.-изд. л.)

40. Mikhailov D. V. Formation and clustering of Russian's nouns's contexts within the frameworks of splintered values / D. V. Mikhailov, G. M. Emelyanov // 9th Int. Conf. on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-9-2008). Conf. Proc. Nizhni Novgorod: N.I. Lobachevsky State University of Nizhni Novgorod, 2008. Vol. 2. P. 39–42. 0,23 п.л. (вклад автора – 0,16 п.л.)

41. Mikhailov D. V. Forming and clustering of syntactic relations on the bases of Natural Language's using's situations / D. V. Mikhailov, G. M. Emelyanov // Interactive Systems And Technologies: The Problems of Human-Computer Interaction. Collections of scientific papers. Ulyanovsk: ULSTU, 2009. Vol. III. P. 295–307. 1,05 п.л. (вклад автора – 0,94 п.л.)

42. Михайлов Д. В. Автоматизация накопления знаний о синонимии текстов предметного языка / Д. В. Михайлов, Г. М. Емельянов // Распознавание-2010: сб. мат-лов IX Междунар. конф. Курск: Курский гос. техн. ун-т, 2010. С. 186–188. 0,12 п.л. (вклад автора – 0,11 п.л.)

43. Михайлов Д. В. Семантическая схожесть текстов в задаче автоматизированного контроля знаний / Д. В. Михайлов, Г. М. Емельянов // 8-я Междунар. конф. "Интеллектуализация обработки информации" (ИОИ-2010): Сб. докл. М.: Макс Пресс, 2010. С. 516–519. 0,50 п.л. (вклад автора – 0,45 п.л.)

44. Mikhailov D. V. Semantic clustering in a problem of text information's compression / D. V. Mikhailov, G. M. Emelyanov // 10th Int. Conf. on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-10-2010). Conf. Proc. St. Petersburg: Politechnika, 2010. Vol. 2. P. 193–196. 0,47 п.л. (вклад автора – 0,31 п.л.)

Доклады на всероссийских конференциях

45. Емельянов Г. М. Вопросы моделирования семантической связанности для систем автоматизированного тестирования знаний / Г. М. Емельянов, Д. В. Михайлов // Всерос. конф. ММРО-10. М.: АЛЕВ-В, 2001. С. 53–56. 0,19 п.л. (вклад автора – 0,17 п.л.)

46. Емельянов Г. М. Вопросы построения механизма суммирования смысла для систем распознавания текстов на естественном языке / Г. М. Емельянов, Д. В. Михайлов // Методы и средства обработки сложной графической информации: тез. докл. VI Всерос. конф. с участием стран СНГ; Нижний Новгород: НИИ прикладной математики и кибернетики ННГУ, 2001. С. 83–85. 0,23 п.л. (вклад автора – 0,21 п.л.)

47. Михайлов Д. В. Пополнение словаря моделей управления в задаче анализа семантической эквивалентности текстовых документов / Д. В. Михайлов, Г. М. Емельянов // Методы и средства обработки сложной графической

информации: тез. докл. VIII Всерос. науч. конф. Нижний Новгород: ГНУ “НИИ ПМК ННГУ”, 2005. С. 88–93. 0,48 п.л. (вклад автора – 0,44 п.л.)

48. Михайлов Д. В. Применение семантических полей словаря РОСС в задаче построения модели управления предикатного слова / Д. В. Михайлов, Г. М. Емельянов // Всерос. конф. ММРО-12. М.: Макс Пресс, 2005. С. 382–385. 0,19 п.л. (вклад автора – 0,17 п.л.)

49. Михайлов Д. В. Кластеризация семантических знаний в задаче распознавания ситуаций смысловой эквивалентности / Д. В. Михайлов, Г. М. Емельянов // Всерос. конф. ММРО-13. М.: Макс Пресс, 2007. С. 500–503. 0,19 п.л. (вклад автора – 0,17 п.л.)

50. Михайлов Д. В. Морфология и синтаксис в задаче семантической кластеризации / Д. В. Михайлов, Г. М. Емельянов // Всерос. конф. ММРО-14. М.: Макс Пресс, 2009. С. 563–566. 0,47 п.л. (вклад автора – 0,42 п.л.)

Свидетельство об официальной регистрации программы для ЭВМ

51. Свидетельство об официальной регистрации прогр. для ЭВМ № 2010617263. Программа формирования синтаксических отношений на множестве семантически эквивалентных фраз / Залешин М. В., Михайлов Д. В., Емельянов Г. М. // заявитель и правообладатель “Новгородский государственный университет имени Ярослава Мудрого”. Заявка № 2010615398; заявл. 02.09.10.; зарег. 29.10.10. 0,10 п.л. (вклад автора – 0,05 п.л.)

Учебники и учебные пособия

52. Михайлов Д.В. Функциональное и логическое программирование. Часть 1. Функциональное программирование. Лабораторный практикум / Д.В. Михайлов, Г.М. Емельянов. Великий Новгород: НовГУ им. Ярослава Мудрого, 2007. 80 с. 4,5 п.л. (вклад автора – 4,1 п.л.)

53. Михайлов Д.В. Функциональное и логическое программирование. Часть 2. Логическое программирование. Лабораторный практикум / Д.В. Михайлов, Г.М. Емельянов. Великий Новгород: НовГУ им. Ярослава Мудрого, 2008. 105 с. 5,9 п.л. (вклад автора – 4,9 п.л.)

Наиболее значимые публикации в других изданиях

54. Емельянов Г.М. Построение динамической модели естественного языка применительно к разработке языковой базы знаний / Г.М. Емельянов, Е.И. Зайцева, Д.В. Михайлов // Искусственный интеллект. 2002. № 2. С. 443–446. 0,35 п.л. (вклад автора – 0,21 п.л.)

55. Емельянов Г. М. Установление смысловой эквивалентности высказываний: на пути к решению проблемы / Г. М. Емельянов, Д. В. Михайлов // Искусственный интеллект. 2004. № 2. С. 86–90. 0,44 п.л. (вклад автора – 0,39 п.л.)

56. Емельянов Г. М. Построение модели управления предикатного слова на основе его лексикографического толкования / Г. М. Емельянов, Д. В.

Михайлов // Таврический вестник информатики и математики. 2005. № 1. С. 35–48. 1,03 п.л. (вклад автора – 0,93 п.л.)

57. Михайлов Д. В. Модель сортовой системы языка в задаче построения семантического образа высказывания на уровне глубинного синтаксиса / Д. В. Михайлов, Г. М. Емельянов // Таврический вестник информатики и математики. 2006. № 1. С. 79–90. 0,56 п.л. (вклад автора – 0,50 п.л.)

58. Емельянов Г. М. Концептуально-ситуационное моделирование процесса перифразирования высказываний естественного языка как обучение на основе прецедентов / Г. М. Емельянов, А. Н. Корнышов, Д. В. Михайлов // Искусственный интеллект. 2006. № 2. С. 72–75. 0,35 п.л. (вклад автора – 0,06 п.л.)

59. Михайлов Д. В. Формирование и кластеризация понятий на основе множества ситуационных контекстов / Д. В. Михайлов, Г. М. Емельянов, Н. А. Степанова // Таврический вестник информатики и математики. 2008. № 2. С. 79–88. 0,32 п.л. (вклад автора – 0,26 п.л.)

60. Емельянов Г. М. Применение реляционной модели представления данных для организации словаря в системе анализа семантической эквивалентности текстов естественного языка [Электронный ресурс] / Г. М. Емельянов, Д. В. Михайлов, Д. В. Силанов // Ученые записки НовГУ. Режим доступа: http://admin.novsu.ac.ru/uni/scpapers.nsf/publications (дата обращения: 23.06.2011). 0,58 п.л. (вклад автора – 0,23 п.л.)

61. Михайлов Д. В. К вопросу автоматизации пополнения базы данных лексических функций в задаче установления смысловой эквивалентности текстов естественного языка / Д. В. Михайлов, Г. М. Емельянов // Вестник Новгородского гос. ун-та им. Ярослава Мудрого, сер. “Технические науки”. 2007. № 44. С. 45–49. 0,47 п.л. (вклад автора – 0,42 п.л.)

62. Михайлов Д. В. Формирование и кластеризация контекстов для существительных русского языка в рамках конверсивных замен / Д. В. Михайлов, Н. А. Степанова, И. И. Юрченко // Физика и механика материалов: прил. к науч.-теорет. и прикл. журн. “Вестник Новгородского государственного университета имени Ярослава Мудрого”. 2009. № 50. С. 31–34. 0,47 п.л. (вклад автора – 0,29 п.л.)

63. Михайлов Д. В. Формирование и кластеризация знаний о синонимии в рамках стандартных лексических функций / Д. В. Михайлов, Г. М. Емельянов // Сб. науч. статей; НовГУ им. Ярослава Мудрого. Великий Новгород, 2009. С. 17–33. 0,99 п.л. (вклад автора – 0,90 п.л.)

Список литературы

1. Аванесов В. С. Композиция тестовых заданий: учебная книга для преподавателей вузов, учителей школ, аспирантов и студентов педвузов / В. С. Аванесов. М.: Адепт, 1998. 217 с.

2. Красильникова В. А. Подготовка заданий для компьютерного тестирования: метод. рекомендации / В. А. Красильникова. Оренбург: ИПК ГОУ ОГУ, 2004. 31 с.

3. Майоров А. Н. Теория и практика создания тестов для системы образования. Как выбирать, создавать и использовать тесты для целей образования / А. Н. Майоров. М.: Нар. образование, 2000. 351 с.

4. Emelyanov G. M. Tree Grammars in the Problems of Searching for Images by Their Verbal Descriptions / G. M. Emelyanov, T. V. Krechetova, E. P. Kurashova // Pattern Recognition and Image Analysis. 2000. Vol. 10. N 4. P. 520–526.

5. Гладкий А. В. Грамматики деревьев. I. Опыт формализации преобразований синтаксических структур естественного языка / А. В. Гладкий, И. А. Мельчук // Информационные вопросы семиотики, лингвистики и автоматического перевода. М., 1971. Вып. 1. С. 16–41.

6. Апресян Ю. Д. Избранные труды: в 2 т. Т. 1: Лексическая семантика. Синонимические средства языка / Ю. Д. Апресян. М.: Языки рус. культуры, 1995. 472 с.

7. Мельчук И. А. Опыт теории лингвистических моделей “Смысл⇔Текст”: Семантика, синтаксис / И. А. Мельчук. М.: Языки рус. культуры, 1999. 345 с.

8. Осипов Г. С. Приобретение знаний интеллектуальными системами: основы теории и технологии / Г. С. Осипов. М.: Наука, 1997. 112 с.

9. Borschev Vladimir. Genitives, Types and Sorts: The Russian Genitive of Measure [Электронный ресурс] / Vladimir Borschev, Barbara H. Partee. Режим доступа: http://semanticsarchive.net/Archive/GJlMzYwN/B&P_PossWkshp04.pdf (дата обращения: 25.06.2011).

Изд. лиц. ЛР № 020815 от 21.09.98.
Подписано в печать . . 2009. Бумага офсетная. Формат 60×84 1/16.
Гарнитура Times New Roman. Печать офсетная.
Усл. печ. л. . Уч.-изд. л. . Тираж экз. Заказ №
Издательско-полиграфический центр Новгородского
государственного университета им. Ярослава Мудрого.
173003, Великий Новгород, ул. Б. Санкт-Петербургская, 41.
Отпечатано в ИПЦ НовГУ. 173003, Великий Новгород,
ул. Б. Санкт-Петербургская, 41.