НАУЧНО - ПРОИЗВОДСТВЕННОЕ ОБЪЕДИНЕНИЕ
"ГОРСИСТЕМОТЕХНИКА"
при *КИЕВСКОМ ГОРИСПОЛКОМЕ*

На правах рукописи

РУБАШКИН Валерий Шлемович

ПРЕДСТАВЛЕНИЕ И АНАЛИЗ СМЫСЛА В ИНТЕЛЛЕКТУАЛЬНЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Специальность: 05.13.17
Теоретические основы информатики

ДИССЕРТАЦИЯ

на соискание ученой степени доктора технических наук

В форме научного доклада

Санкт-Петербург
1992

Работа выполнена
Петербургского институт

Официальные оппоненты:

- доктор технических наук, профессор
Поспелов Д.А. (г. Москва)

- доктор технических наук, профессор
Александров В.В. (г. Санкт-Петербург)

- доктор технических наук, профессор
Довгялло А.М. (г. Киев)

Ведущая организация - Санкт-Петербургский государственный университет, НИИ математики и механики.

Защита состоится ~~"21 октября"~~ 4 ноября 1992 г.
на заседании Специализированного Совета Д.166.01.01
по присуждению ученой степени доктора наук в НПО
"Горсистемотехника" по адресу:
252004, Киев, ул. Красноармейская, 23б, конференц-зал.

С диссертацией можно ознакомиться
в библиотеке НПО "Горсистемотехника"

Автореферат разослан "2" октября 1992 г.

Ученый секретарь
Специализированного Совета
кандидат технических наук Михновский С.Д.

1. ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Научная проблема. Актуальность исследования.

В настоящее время формируется новое поколение информационных технологий, основанных на концепции интегрированной информационной среды (multimedia), объединяющей инструментарий машинной графики, многооконных систем, гипертекстовых систем, систем управления базами данных, метода "активных зон" и т. д. Одновременно сформировался и стал доступен широкому пользователю огромный поток полнотекстовых баз данных, включающих мировой поток научно-технической литературы, поток текстов, порождаемых средствами массовой коммуникации, потоки деловой документации в текстовой и частично структурированной бланковой форме и т. п. При этом становится очевидной растущая диспропорция между объемом потенциально доступной информации и мощными средствами манипулирования хорошо структурированными данными, с одной стороны, и традиционными ("ручными", "интеллектуальными") методами систематизации и отбора, вообще структурирования текстов на естественном языке, с другой стороны. Важными направлениями такого структурирования являются, например, формирование фактографических баз данных по входному потоку текстов и формирование гипертекстовых структур.

Как одно из основных направлений преодоления указанной диспропорции следует рассматривать усилия по разработке информационно-лингвистических процессоров (ИЛП) - класса лингвистических процессоров, ориентированных на глубокую смысловую обработку делового текста. Речь идет о лингвистическом процессоре "логического" типа, результатом работы которого является адекватное представление содержащихся в тексте знаний в форме, обеспечивающей далее реализацию механизмов логического вывода. Такой процессор противопоставляется процессору "трансферного" типа, результатом работы которого является установление синтаксической структуры предложений анализируемого текста - с ориентацией, главным образом, на задачу автоматического перевода.

Краткая характеристика состояния проблемы.

Проблема разработки средств для формализации когнитивного содержания текста изучалась в рамках двух дисциплинарных традиций. Одна из них - это прикладная лингвистика, другая - документальные информационно-поисковые системы (ДИПС). Первая была и остается ориентированной преимущественно на задачу автоматического перевода, вторая - на проблему поиска научно-технического (или шире - делового) документа путем анализа - разумеется, частичного и приближенного - его содержания. Разработки и исследования ДИПС с самого начала были - по самой постановке проблемы - ориентированы на представление и анализ именно смысловой стороны сообщений. Кроме того, в русле ДИПС по существу впервые были предприняты широкие работы по практическому моделированию терминосистем в интересах задач автоматического анализа делового текста (разработка тезаурусов).

Интерес к семантическим аспектам анализа текста возникал, как правило, на пересечении этих двух направлений. Пионерскими - с точки зрения их интеграции - были работы группы Э. Ф. Скороходько в Киеве в начале 60-х гг. Позже (в конце 60-х - начале 70-х гг.) они дали начало более узкому инженерному направлению в моделировании смысла, получившему название "ситуационное управление" (Ю. И. Клыков, Д. А. Поспелов). Характерными для данного подхода можно считать моделирование смысла текста в языке бинарных отношений - триад вида R(a,b). Для представления связей между лексическими единицами в словаре или в тексте может быть использована также сетевая нотация, в которой бинарному отношению соответствует дуга семантической сети, соединяющая узлы a и b.

С конца 60-х гг. другой вариант триадной схемы представления смысла развивался Н. Н. Леонтьевой. Отличительная особенность ее подхода - преимущественное внимание к семантическому моделированию предикатно-актантных связей ("связей по валентности"). Н. Н. Леонтьевой был предложен один из первых вариантов "смысловой грамматики", включающей набор "элементарных смысловых отношений"; ею предприняты так-

же едва ли не первые попытки дать семантическое описание таких лексических систем русского языка как предлоги и слова со значением времени.

Различные варианты представления смысла делового текста на языке семантических графов были предложены в начале 80-х гг. в работах Г. С. Цейтина и Э. В. Попова. При этом Г. С. Цейтин исходил из задачи интегрировать в едином языке как собственно средства представления знаний, так и средства описания процедур анализа. Получившийся формализм был назван "ассоциативной сетью". А для работ Э. В. Попова отличительной чертой можно считать стремление дополнить модели анализа текста "моделью участников общения".

Интересные работы по созданию информационно-лингвистического процессора, ориентированного на формализацию когнитивного содержания научно-технического текста, ведутся группой под руководством М. Г. Мальковского. Здесь, в частности, развиты интересные методы адаптации словарей системы - в том числе и семантических - к новым предметным областям. Группой под руководством Г. Г. Белоногова сформированы уникальные по объему словари и развиты методы упрощенного анализа больших политематических текстовых потоков. В последние годы определенное внимание семантическому компоненту лингвопроцессора уделяется в исследованиях коллектива, руководимого Ю. Д. Апресяном. Однако эти работы ограничены пока достаточно узкой задачей перевода запросов к реляционной базе данных с естественного языка на язык запросов типа SQL. Основные результаты, полученные указанными исследовательскими группами, широко известны в профессиональной среде.

Исследования по данному кругу проблем в Западной Европе и США имеют более прочную традицию и ведутся широким фронтом. Показателен в этом плане, например, тот факт, что до 3/4 всех публикаций, помещаемых в журнале Computational Linguistic в последние несколько лет, посвящены проблемам семантики текста. Ежегодно публикуются десятки монографий, фактически формирующих новую дисциплинарную традицию. К сожалению, ничего аналогичного этому широкому научному движе-

нию ни в России, ни на Украине не наблюдается. Таким образом, обозначилось серьезное отставание еще в одной весьма перспективной области исследований. Достаточно полное представление о состоянии зарубежных исследований и разработок в этой области, их идейно-теоретической базе могут дать монография Graeme Hirst. Semantic interpretation and the resolution of ambiguity. - Cambridge University Press, 1987 и текущее содержание названного выше журнала. Один из существенных уроков, который можно извлечь при изучении состояния исследований на Западе, состоит в том, что оказался практически несостоятельным пуритански-теоретический подход к формализации семантики текста, ассоциировавшийся в значительной степени с работами Монтегю.

Несмотря на большие усилия, предпринимаемые научным сообществом - особенно в последние годы, приходится вслед за большинством специалистов констатировать, что на сегодня отсутствует общепризнанная методология построения послесинтаксического этапа анализа делового текста. Модели и методы, используемые при разработке информационно-лингвистических процессоров, пока фрагментарны и не образуют целостной теоретически обоснованной концепции. Определились наиболее сложные проблемы, решение которых требует привлечения принципиально новых идей:

- методы использования знаний о предметной области в процедурах анализа текста;

- методы формализации и представления знаний, пригодные для использования в лингвистических процессорах;

- методы разрешения морфологических, синтаксических и лексических неоднозначностей.

В настоящей работе именно эти вопросы находились в центре внимания и, как нам представляется, по этим направлениям в ней получены заслуживающие внимания результаты.

Круг идей, представленный в настоящей работе, имеет своим источником работу автора в двух внешне далеких друг от друга областях: логический анализ языка науки ([1], [2], [3], [10]) и документальные информационно-поисковые системы([4], [8], [9]). В первом направлении автора интересовал язык неклассической физики - точнее, возможность его логической экспликации и прояснения на этом пути парадоксальных, с точки зрения здравого смысла и классической науки, положений квантовой теории и теории относительности. Содержанием работы во втором направлении был поиск путей совершенствования информационных языков с целью улучшения технических характеристик документальных ИПС. И хотя это может показаться странным, из соединения того и другого возник комплекс идей, определивших новый подход к проблеме автоматического анализа делового текста.

Основная цель исследования.

Определить и практически проверить научно-методические и инженерные основы разработки лингвистических процессоров нового типа.

Научная новизна результатов исследования.

В ходе работы по решению указанной выше общей проблемы были получены следующие конкретные научные результаты.

1. Разработаны оригинальные модели и методы информационного анализа и формализации делового текста, основанные на использовании знаний о предметной области и ориентированные на применение в современных информационных технологиях, в том числе:

1.1. Предложен и теоретически обоснован язык представления информационного содержания делового текста, представляющий собой конкретный вариант языка семантических сетей, допускающий простой переход от сетевой нотации к логической и обратно.

1.2. Разработаны модели и методы концептуальной интерпретации синтаксических структур, обеспечивающие:
- определение семантических ролей объектов в предложении;
- анализ специальных видов синтаксической связи;
- лексикализацию синтаксических связей на основании знаний о предметной области;
- разрешение лексической и синтаксической омонимии;
- интерпретацию слабых синтаксических связей.

1.3. Разработаны модели и методы установления межфразовых связей (кореференция имен).

2. Разработаны модели и методы формализации априорных знаний о предметной области, используемых в процедурах анализа.

3. Разработаны и практически проверены методы программной реализации предложенных моделей.

Апробация работы.

Основное содержание работы отражено в монографии "Представление и анализ смысла в интеллектуальных информационных системах".

Всего по теме диссертации опубликовано свыше 50 работ; 24 работы, достаточно полно отражающие содержание выполненных исследований, указаны в настоящем докладе.

Полученные в ходе исследований научные результаты - по мере их наработки - докладывались и обсуждались на конференциях и симпозиумах по прикладной лингвистике, информатике, логике и методологии науки (свыше 20 докладов).

Результаты исследования были положены в основу разработки двух программных систем:
- прототипная система информационного анализа научно-технического текста для ЕС ЭВМ (1983-1987 г.);
- семантический компонент информационно-лингвистического процессора для персональных ЭВМ (1989-1992 г.).

Теоретические и прикладные результаты, полученные в ходе исследований, были использованы при выполнении плановых НИР и ОКР во ВНИИ "Информэлектро" и НПО "Персей".

2. МОДЕЛИ И МЕТОДЫ КОНЦЕПТУАЛЬНОГО АНАЛИЗА ДЕЛОВОГО ТЕКСТА

2.1. Замечания по постановке задачи ([18], [24]).

Наиболее существенными чертами систем рассматриваемого типа можно считать следующие:

- допустимость частичного анализа и постредактирования (в широком смысле);
- объектографическая ориентированность анализа;
- открытые словари, ориентированные на достаточно широкую предметную область;
- сравнительно высокий уровень моделирования как языковой, так и профессиональной компетенции.

С учетом содержательно-стилистических особенностей научно-технических текстов (в частности, - текстов, относящихся к предметной области "Машиностроение") и класса решаемых информационных задач оправданы следующие ограничения по типам формализуемых структур содержания.

Имеют приоритет в анализе следующие типы содержания:

- имена объектов, описываемых в тексте, и связи кореференции между ними, включая и межфразовые связи этого типа;
- имена признаков и их связи с именами объектов;
- имена процессов/действий и их связи с объектами;
- имена отношений, значимых в данной предметной области (ограниченный список), и их связи с именами объектов.

Как правило, игнорируются следующие типы содержания:

- сюжетные связи между процессами/действиями (последовательность во времени, причинные и целевые связи и т. п.);
- модальность пропозиций и все элементы содержания, выраженные грамматическими средствами (число, соотношение грамматических времен, наклонение и т. п.)
- метатекстовые элементы (характеризующие последовательность изложения, оценки излагаемых фактов, мотивировки и аргументация и т. п.).

Игнорируется кванторная информация. Принимается допущение, согласно которому все входные сообщения рассматриваются как "факты".

2.2. Общая схема анализа текста ([15], [23], [24]).

Информационно-лингвистический процессор реализует все традиционно рассматриваемые в прикладной лингвистике и информатике уровни анализа текста: морфологический, синтаксический, семантический, анализ дискурса.

Собственно языковое содержание процедур анализа может быть охарактеризовано следующим образом.

1. Морфологический анализ.

Поддерживается системой словарей лингвистического уровня. Включает процедуру грамматической квалификации слов, отсутствующих в словаре основ.

Выход: последовательность лексем с указанием грамматических характеристик представляющей лексему словоформы и всех ее понятийных эквивалентов (допускается грамматическая и семантическая неоднозначность).

2. Синтаксический анализ.

Выход: дерево зависимостей (все полученные варианты) с указанием лексических (понятийных) вариантов. Специальным "нулевым дескриптором" отмечается отсутствие лексемы в позиции сильноуправляемого слова.

3. Терминологический и признаковый анализ.

Распознаются терминологические словосочетания, заданные в словаре словосочетаний в виде синтаксического графа. Распознаются также конструкции типа "числовой признак с вещественным значением", "целочисленный признак", "символьный признак".

Выход: подграф, представляющий в синтаксическом графе каждую из названных конструкций, заменяется одним узлом, с сохранением за ним грамматического статуса главного слова заменяемой конструкции. (Семантические характеристики узла при этом изменяются.)

4. Локальный семантический анализ по предложениям (концептуальная интерпретация синтаксического графа предложения).

Подчинительные связи в дереве синтаксических зависимостей просматриваются в порядке снизу вверх, слева направо и интерпретируются в терминах семантического представления.

Выход: семантический граф предложения (возможно, несвязный).

5. Анализ межфразовых связей.

Устанавливаются отношения референциального тождества имен объектов по всему тексту. Отождествляются имена процессов, отношений, признаков, устраняется смысловая избыточность.

Выход: семантический граф, бесповторно отображающий основное информационное содержание связного текста.

2.3. Интерпретация связей внутри именных групп с предметным значением ([6], [7], [21) [23].).

Связи этого типа - вместе с некоторыми другими - именуются обычно слабыми связями.

Для связей данного типа допускаются следующие варианты интерпретации:

1. Установление кореференции (синтаксические узлы отца и сына склеиваются в один семантический узел). Например, сочетание "медная пластина" трансформируется в дескрипцию вида

МЕДЬ(x) & ПЛАСТИНА (x).

Основанием для установления кореференции является совместимость термов, устанавливаемая обращением к базе знаний (гипотеза локальной кореференции).

2. Установление определительного (неспецифицированного) отношения: синтаксические отец и сын соединяются дугой OF. Так, например, может быть интерпретировано сочетание *ротор двигателя* при отсутствии дополнительных сведений об упоминаемых здесь реалиях в базе знаний.

3. Установление отношения "объект - назначение". Синтаксические отец и сын соединяются дугой FN : *сварочное оборудование*.

4. Лексикализация синтаксической связи (реализовано для связей типа "часть - целое") - основывается на знаниях о предметной области.

Лексикализация отношения производится одним из двух способов:

- поиском в базе знаний отношения, ассоциированного с парой (хозяин, слуга);
- в случае предложной связи - поиском в базе знаний отношения, присоединенного к предлогу.

Так, например, сочетание *ротор двигателя* трансформируется в дескрипцию вида

РОТОР(y) & ДВИГАТЕЛЬ(x) & ИМЕТЬ_ЧАСТЬЮ(x, y) ,

в случае, если указанное ассоциированное отношение для данной пары (с учетом наследования свойств) в базе знаний найдено (см. рис. 1.).

Определение типа интерпретации выполняется методом разбора случаев с существенным использованием базы знаний.

2.4. Модели и методы анализа предикатно-актантных связей ([6], [21], [23]).

Здесь решались следующие основные вопросы.

1. Определение необходимого и достаточного набора семантических ролей (номенклатура валентностей).
2. Способ описания моделей управления у предикатных термов.
3. Способ установления соответствия между грамматической ролью имени в предложении и его семантической ролью.

В основу описания семантических моделей управления положены две следующие рабочие гипотезы, вполне подтвердившиеся при работе с реальным языковым материалом.

Гипотеза 1. Для выражения основного информационного содержания научно-технического текста достаточен следующий минимальный набор имен валентностей:

OB, OB1, OB2, INS, SB1, SB2.

Типовое употребление имен валентностей:

OB

- объект, на который направлено действие, субстанция процесса (*нагрев двигателя, рост кристаллов, расчет магнитопровода, регулирование скорости, сборка ротора, импорт нефти);*
- нейтральный объект тернарного отношения *(между крышкой и корпусом монтируется прокладка);*
- признак - носитель признака *(надежность двигателя, диаметр вала).*

Text Tracing Analysis

Text

Вал вращается асинхронным двигателем с фазным ротором мощностью 400 вт

ИМЕТЬ_ЧАСТЬЮ <Esc>

Semantic Interpretation

Father	sct	Arc	Son	sct
1. АСИНХРОННЫЙ_ДВИГАТЕЛЬ	3.2	ASR--->1.	ФАЗНЫЙ_РОТОР	3.1
Syntax Variant of Father:				
1. ВРАЩАТЬ	6.2	1.	ФАЗНЫЙ_РОТОР	3.1

Рис. 1. Пример лексикализации синтаксической связи с одновременным разрешением синтаксической омонимии.

OB1, OB2
- актанты бинарного отношения *(датчик установлен на крышке)*;
- актанты действия, отношения, характеризующегося направленностью от начального пункта/состояния к конечному *(кремний осаждается из паровой фазы, импорт нефти из Венесуэлы в США, давление зависит от температуры)*;

INS
- инструмент, средство, метод, агент, используемые для достижения цели или участвующие в процессе *(электроискровая обработка металлов, тиристорное управление двигателем, охлаждение с помощью тепловых труб)*.

SB1
- активный (действующий) субъект;

SB2
- пассивный субъект ("пациенс").

Гипотеза 2. Словарь предикатных термов может быть описан конечным, и притом, обозримым списком моделей управления (несколько десятков моделей). Практически возможно разбить словарь предикатных термов на содержательные классы, соотносимые с определенным типом модели управления.

Модели управления приписываются только термам семантических категорий "ПРОЦЕССЫ/ДЕЙСТВИЯ" и "СТАТИЧЕСКИЕ ОТНОШЕНИЯ". Наличие модели управления у термов указанных классов не является обязательным.

Для семантической модели управления (СЕМУ) принята следующая формализация.

```
СЕМУ ::= НОМЕР_СЕМУ <описатель валентности> |
         <СЕМУ> <описатель валентности>
<описатель валентности> ::= <имя валентности> <семантическое
                            условие заполнения> <облигат>
<имя валентности> ::= OB1 | OB2 | OB | INS | SB1| SB2
<семантическое условие заполнения> ::= <категориальный ограничи
                                       тель> < конкретизатор>
<категориальный ограничитель> ::= ЛОГИЧЕСКАЯ_ШКАЛА |
```

<конкретизатор> ::= КОД_ДЕСКРИПТОРА | ПУСТО
<облигат> ::= + | -
Здесь:

- Разрядность ЛОГИЧЕСКОЙ_ШКАЛЫ_1 определяется числом семантических категорий, различаемых в словаре (семь); установка бита, соответствующего номеру семантической категории, определяет допустимость заполнения валентности термином данной категории.

- Код дескриптора в позиции "конкретизатор" указывает, что валентность может быть заполнена только дескриптором, совместимым с данным. Конкретизатор может быть пуст (= 0).

- Облигат определяет обязательность(+) или факультативность(-) заполнения валентности.

Процедура заполнения валентностей предикатного терма в тексте алгоритмически реализована как функция, устанавливающая соответствие между грамматической формой предиката, его семантическим типом и грамматической ролью актанта, с одной стороны, и семантической ролью актанта, с другой. Функция задается в форме таблицы что, соответственно, дает возможность легко изменять ее при настройке на конкретный язык.

2.5. Модели и методы разрешения лексической и синтаксической омонимии ([23]).

Предлагаемая техника разрешения лексической омонимии состоит в следующем. Все возможные типы семантической интерпретации ранжируются по степени успешности. Ранг интерпретации определяется приписываемой ей весовой оценкой. В процессе интерпретации текущей синтаксической связи последовательно просматриваются все лексические варианты сына, все лексические варианты отца. Каждый из получаемых вариантов интерпретации сравнивается по степени успешности с предыдущими вариантами интерпретации текущей синтаксической связи. Если полученный вариант хуже предыдущих - он исключается; если лучше - исключаются все предшествующие интерпретации; если равен предыдущим - вариант присоединяется к предыдущим, т.е. неоднозначность сохраняется для дальнейшего анализа.

Пример успешного разрешения лексической омонимии по синтаксическому контексту показан на рис. 2. Анализируется пара *ремонт установок*. Слово *установка* обнаруживает возможность двух понятийных интерпретаций: (1) 'техническое устройство'; (2) 'устанавливать'. Интерпретатор выбирает в качестве правильного первое из указанных значений.

В основу алгоритма разрешения синтаксической омонимии положен тот же механизм ранжирования интерпретаций. Однако в этом случае ранжируются оценки интерпретации для предложения в целом. Имеется в виду, что на вход семантического интерпретатора поступают последовательно все варианты синтаксического разбора предложения. Для каждого варианта вычисляется суммарная оценка качества интерпретации, определяемая как сумма весов интерпретаций всех парных связей. При обнаружении незаполненных обязательных валентностей их вес вычитается из суммарной оценки. Для сочинительной связи оценивается степень семантической согласованности сочиненных элементов. В результате выбирается вариант (варианты) разбора, получившие максимальную оценку. Пример разрешения синтаксической омонимии показан на рис. 3. Здесь обнаруживается неоднозначность синтаксического подчинения слова *ротор*. Интерпретатор выбирает вариант *асинхронный двигатель мощностью ...* .

2.6. Модели и методы установления межфразовых связей ([13], [14], [15], [21], [23]).

С точки зрения принимаемой постановки задачи основной проблемой при анализе межфразовых связей является установление кореференции имен объектов. К этому моменту текст представлен в виде несвязного семантического графа, в котором отдельные связные фрагменты отражают результаты семантико-синтаксического анализа по предложениям. В процессе такого анализа могут быть распознаны имена ситуаций ("предикаты"), их актанты, могут быть различены и отображены в локальной семантической структуре объекты внутри именных групп. Можно считать, что в таком графе сохранена также позиционная информация: например, для каждого узла указана позиция в тексте последнего термина, отнесенного к данному узлу.

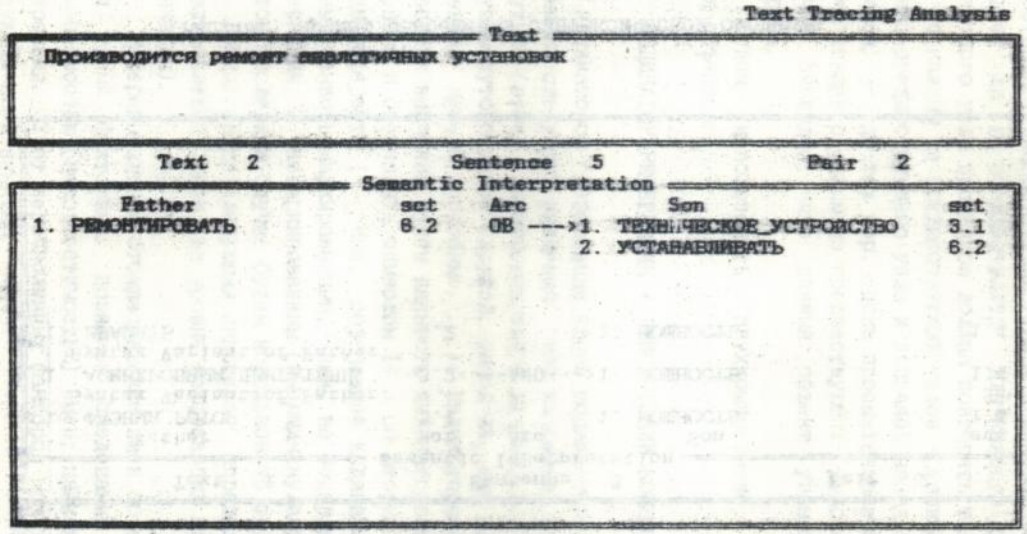

Рис. 2. Пример разрешения лексической омонимии.

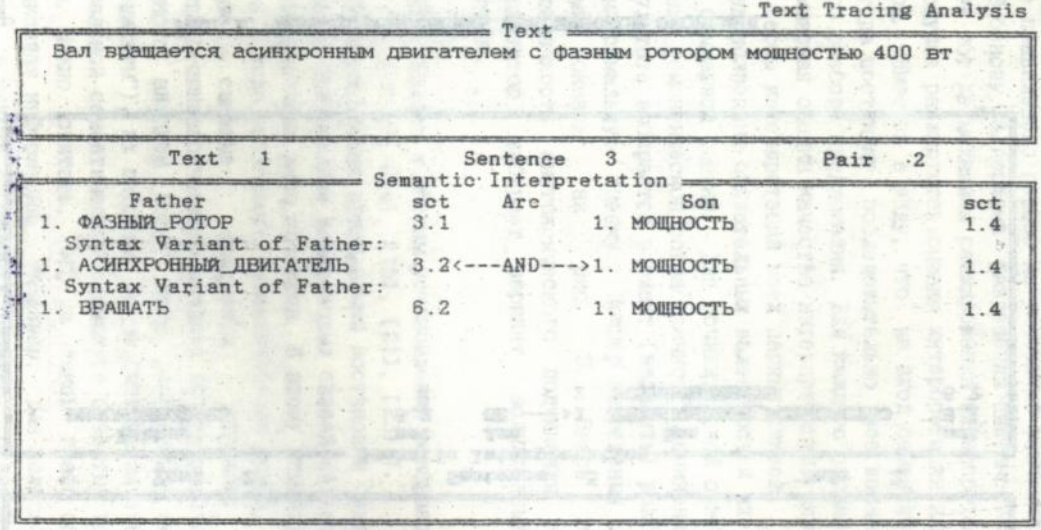

Рис. 3. Пример разрешения синтаксической омонимии.

Далее должен решаться вопрос о способе объединения названных фрагментов графа - путем референциального отождествления упоминаемых в тексте объектов и склеивания соответствующих им узлов. Такой анализ можно представить себе как многократное повторение одной и той же процедуры, которая сопоставляет два объектных узла и принимает решение о том, следует ли их отождествлять. Для каждого просматриваемого объектного узла процедура должна повторять такое сопоставление с каждым из предшествующих узлов - начиная с непосредственно предшествующего узла и двигаясь назад к началу текста (абзаца) - до тех пор, пока просматриваемый узел не будет отождествлен с одним из предшествующих узлов. В противном случае очевидно, что данный объект упоминается в тексте впервые.

Будем использовать двухкомпонентное представление имен объектов:

ИНДИКАТОР РЕФЕРЕНЦИИ + ЛЕКСИЧЕСКИЙ КОМПОНЕНТ.

Под индикатором референции здесь понимается любой элемент с одним из следующих значений: 'тот же самый, что и ранее упомянутый' (*этот, вышеупомянутый, данный, ...*), 'отличный от упомянутого ранее'(*другой, такой же, еще один, остальные, прочие, подобные, ...*). Индикаторы с первым из указанных значений будем называть индикаторами тождества референтов и обозначать символом I+, со вторым - индикаторами смены референта I-. Имя, содержащее индикатор референции, будем называть маркированным; имя, не содержащее индикатора референции, - немаркированным (имеющим нулевую маркировку). В этом последнем случае будем использовать символ I0.

Предлагается следующая гипотеза, определяющая механизм отождествления имен на уровне межфразовых связей (гипотеза индикации).

Референциальное отождествление имен объектов определяется тремя факторами: порядком следования имен в тексте; совместимостью (несовместимостью) имен; наличием индикаторов референции. Учет перечисленных факторов может быть сведен к следующим двум пунктам.

1. Несовместимость имен является достаточным условием их референциального различия; при следовании друг за другом несовместимых имен смена референта не маркируется. (Этим, в частности, можно объяснить аномальность текстов такого рода: *Завод "Электросила" выпускает крупные электрические машины. Другие трансформаторы заводом не производятся.* Ср. вариант нормального продолжения: ... *Другие изделия заводом не производятся.*)

2. Совместимые имена по умолчанию (т.е. при отсутствии индикатора смены референта) считаются референциально тождественными. Поэтому маркировка референциального различия для следующих друг за другом совместимых имен является обязательной. Ср., например: *Завод "Электросила" выпускает крупные электрические машины. Аналогичное предприятие находится в Харькове.* Опущение индикатора смены референта (аналогичный) влечет принципиальное изменение смысла: *Завод "Электросила" находится в Харькове.* Маркировка тождества референтов возможна, но не обязательна.

Коротко смысл гипотезы индикации может быть передан следующей формулировкой: для несовместимых имен нулевой индикатор маркирует референциальное различие, для совместимых - референциальное тождество.

В соответствии с принятыми выше допущениями предполагается, что локальные показатели нетождественности (сопредикатность имен, внутренняя структура именных групп и др.) учтены на предшествующем этапе семантико-синтаксического анализа текста по предложениям.

Содержание гипотезы индикации весьма компактно может быть представлено в табличной форме. Таблица 1 отражает точку зрения анализа текста (на входе - сведения о маркированности второго имени и о совместимости имен, на выходе - решение о необходимости референциального отождествления имен). Символы "==" (" ^=") в таблице означают, что при данной комбинации условий имена должны получить один и тот же (разные) референциальные индексы.

Таблица 1

Индикатор референции :	I+	IO	I-
Совместимость имен:			
ДА	==	==	^=
НЕТ	^=	^=	^=

Гипотезу индикации можно рассматривать как межфразовый аналог гипотезы проективности. Действительно, гипотеза проективности, действие которой ограничено пределами предложения, утверждает, что между двумя словами A и B, связанными прямой синтаксической связью, не может находиться слово C, связанное прямой синтаксической связью со словом, находящимся слева от A или справа от B. Аналогичным образом гипотеза индикации утверждает, что в связном тексте между именами A и B, находящимися в отношении референциального тождества, не может находиться имя C, совместимое хотя бы с одним из них (с учетом индикаторов референции) и не находящееся с ними в том же отношении.

В связи с этой аналогией полезно вспомнить, что хотя в реальных текстах условие проективности выполняется далеко не всегда, оно используется как эффективный инструмент синтаксического разбора предложения во всех без исключения синтаксических анализаторах.

Если принять гипотезу индикации, то основные вычислительные сложности при анализе межфразовой кореференции оказываются связаны с вычислением совместимости имен. Здесь существенны лексическое наполнение и контекстное окружение сравниваемых узлов семантической сети.

При сопоставлении узлов, содержащих более одного термина, должны попарно сравниваться на совместимость каждый термин одного узла с каждым из терминов другого. При этом, например, будет установлена совместимость имен *мощные электрические машины постоянного тока* и *электродвигатели на напряжение 6 кВ с номинальной скоростью 900 об/мин* и несовместимость первого из них с именем *3-фазные асинхронные двигатели* (последнее вытекает из несовместимости термина *машина постоянного тока* с каждым из терминов *3-фазный* и *асинхронный двигатель*).

Учет внешнего контекста практически сводится к учету отношений между объектами. Для текстов машиностроительной тематики в первую очередь должны быть приняты во внимание отношения "целое - часть" и "объект-функция". Сведения о несовместимости такого рода конструкций должны храниться в базе знаний. Например, словарно может быть задана несовместимость смыслов 'иметь частью короткозамкнутый ротор' и 'иметь частью фазный ротор'. Аналогично для технических устройств должна быть отражена в базе знаний схема 'если назначение различно, то и устройства различны'; отсюда вытекает несовместимость друг с другом таких сочетаний, как *измерительное устройство, станок для намотки, осветительный прибор* и т. п.

В общем случае установление несовместимости двух имен объектов, которым на семантическом уровне сопоставляются описания D1(x) и D2(x), реализуется выводом по схеме

$$\frac{K \,\&\, D1(x) \longrightarrow A(x),\ K \,\&\, D2(x) \longrightarrow \neg A(x)}{D1(x) \longrightarrow \neg D2(x)}$$

(здесь K - знания, хранящиеся в базе знаний о предметной области, " ¬ " - символ логического отрицания).

3. ЯЗЫКОВЫЕ СРЕДСТВА ПРЕДСТАВЛЕНИЯ ЗНАНИЙ В ИНФОРМАЦИОННО-ЛИНГВИСТИЧЕСКОМ ПРОЦЕССОРЕ

Выбор и обоснование языковых средств ИЛП осуществляется по принципу нисходящего проектирования: движением от универсального логического языка, позволяющего моделировать любые (или, по крайней мере, любые практически значимые) виды языкового содержания, к рабочим языкам ИЛП.

Переход от логического языка к языку ИЛП предполагает следующую последовательность действий:

- определяется базовый логический язык;
- вводятся (и прагматически обосновываются) ограничения на вид допустимых формул;
- вводятся (и прагматически обосновываются) соглашения об умолчаниях;
- вводятся (и прагматически обосновываются) ограничения на используемые средства логического вывода.

При этом нужно ясно понимать, что все упрощения суть неформальные экспертные решения, принимаемые разработчиком на основании имеющегося у него понимания задач и ориентированные на создание эффективной и хорошо сбалансированной системы.

Такой подход позволяет на каждом шаге контролировать "логические потери", вызываемые принятыми упрощениями и, с другой стороны, при необходимости строить логический эквивалент для любой используемой конструкции логического языка.

3.1. Базовый логический язык ([5], [16], [17] [19], [23]).

Идея логического анализа (логического моделирования) языка науки родилась практически одновременно с созданием аппарата современной символической логики. Однако ранние работы 20-х и 30-х годов (работы Карнапа, Рейхенбаха и всей школы "логического эмпиризма") позволили лишь обнаружить глубину проблемы. Сложности, обнаружившиеся при столкновении с реальным языковым материалом, привели к тому, что исследования по приложению аппарата логики к языковым проблемам вне математики в последующие десятилетия были направлены преиму-

щественно на моделирование ограниченных и весьма специализированных понятийных систем: модальная логика, деонтическая логика, логика времени, логика оценок и т. п. Однако основной для практических приложений вопрос - о моделировании некоторого профессионального языка как целого - при этом остался вне рассмотрения. Реальный текст - даже отнесенный к весьма ограниченной предметной области - не может состоять, скажем, только из временных или оценочных терминов. Ясно, что в этом случае требуются существенно более богатые логические модели. Две проблемы, на наш взгляд, вызывали при этом наибольшие трудности. Первая - способ интеграции в едином логическом языке математических и содержательно-понятийных средств представления знаний. Вторая - способ описания семантической сочетаемости лексических единиц (проблема осмысленности). В первом случае речь идет о том, как записать в логическом языке количественные утверждения, соотносящие формульные величины с соответствующими им классами объектов (скажем, записать на логическом языке фразу "*площадь круга (s) равна квадрату его радиуса (r), умноженному на число пи*"). Во втором случае - о том, как логически квалифицировать интуитивную некорректность таких выражений как *синий электрон, синяя температура, синее падение* и т. п.

В настоящей работе описан язык логической систематизации профессиональных знаний (язык ИНФОЛ), предлагающий определенное решение для этих проблем на уровне логического исчисления. В качестве модельной основы такого языка может, видимо, рассматриваться язык теории категорий. Интересная попытка такого рода сравнительно недавно предпринята Е. М. Бениаминовым, предложившим категорный язык представления знаний ЭВОП.

Язык ИНФОЛ дает возможность, в частности:

- задать полную систему объемных отношений между именами объектов;
- логически интерпретировать выражения с числовыми параметрами и собственными именами;
- ввести отношение осмысленности признака для заданного класса объектов, позволяющее установить связь между логическими и фреймовыми языками.

Нотацию, используемую в языке ИНФОЛ, продемонстрируем на упомянутом выше примере. Связь между понятиями "площадь" и "круг", с одной стороны, и соответствующими формульными величинами, с другой, может быть представлена следующим образом. (Пример, в силу своей краткости, чисто иллюстративный.)

```
Ax As Ar ( ( ГЕОМ_ФОРМА(x, КРУГ) & ПЛОЩАДЬ(x, s) & РАДИУС(x, r) )
        ---> s = пи * (r ** 2) )
```

При этом существенно, что аксиоматика, образующая данную теорию (в логическом смысле слова "теория"), должна включать, например, нижеследующие утверждения:

```
Ax As ( ПЛОЩАДЬ(x, s)  --->  ГЕОМ_ТИП(x, ПОВЕРХНОСТЬ) );

   Ax ( ГЕОМ_ТИП(x, ПОВЕРХНОСТЬ)  --->  ^ГЕОМ_ТИП(x, ТЕЛО) );

   Ax ( ВАЛ(x)  --->  ГЕОМ_ТИП(x, ТЕЛО) );
```

Отсюда, в частности, и выводится семантическая некорректность (в данном случае - логическая ложность) выражений типа *площадь куба, площадь вала* и т.п.

3.2. Переход от логического языка к языку представления содержания текста и тезаурусу ([4], [20], [22] [23]).

Прагматические основания для принимаемых упрощений базового логического языка заключены в следующих обстоятельствах.

1. Раздельная организация ввода и хранения фактов и зависимостей.

При этом фактически принимается допущение, согласно которому кванторный префикс с кванторами существования для всех референциальных индексов по умолчанию предшествует целому тексту. Содержательно это означает принятие допущения о том, что входной текст представляет собой совокупность описаний индивидуализированных объектов и связей между ними.

Зависимости же фиксируются в базе априорных знаний о предметной области.

2. Неполнота существующих семантических теорий и, соответственно, ограниченные возможности освоенной в настоящее время техники формализации входных сообщений на естественном языке, равно как и техники формализации рассуждений в достаточно широкой предметной области.

3. Требования эффективности и сбалансированности реализации ИЛП.

В результате разделения в ИЛП фактов и зависимостей базовый язык превращается в два связанных лишь общностью лексики языка: язык описания содержания текста и язык представления универсальных для данной предметной области зависимостей (в терминологии информационных систем - язык индексирования и тезаурус). Каждый из этих языков может иметь свою нотацию, существенно отличающуюся от логической нотации. Основной лексической единицей того и другого языка является дескриптор (в реализации - числовой код понятия). Этой единице может быть поставлена в соответствие логическая формула того или иного вида - в зависимости от категории дескриптора.

Язык индексирования использует всего два типа формул:

объектно-характеристические записи вида

$$F\_1\ (x)\ \&\ F\_2\ (x)\ \&\ \ldots\ \&\ F\_n\ (x)$$

и утверждения о связях вида

$$R\ (x\_1,\ x\_2,\ \ldots\ ,\ x\_m).$$

Соответственно, не требуют явного выражения в языке индексирования логические связки, кванторы, а в объектно-характеристических записях - также и референциальные индексы. С другой стороны, оказывается полезным ввести явное представление для таких элементов, как имя валентности и вид значения; в логической нотации им соответствуют позиции объектных переменных и употребление вспомогательных предикатов. Принятие перечисленных ограничений позволяет перейти в представлении содержания текста от логической нотации к семантической сети.

Тезаурус фиксирует объемные либо ассоциативные (необъемные) отношения между дескрипторами. Здесь отметим только принципиальное различие между собственно зависимостями - все они так или иначе могут быть сведены к схеме формальной импликации Ax (F_1(x) --> F_2(x)) - и формальными толкованиями, соответствующими логической форме определения.

И в языке индексирования и в тезаурусе используются тривиальные схемы вывода:

```
A & B  !--  A
( A --> B ) & ( B --> C )  !--  ( A --> C )
( A --> B ) & ( B --> ^C )  !--  ( A --> ^B)
F & (A --> B)  !--  F[A/B]
```

(F[A/B] обозначает формулу, полученную заменой всех вхождений A на B).

Единственная нетривиальная схема вывода (к тому же используемая эпизодически) имеет вид

```
Ax (F(x) --> ^G(x))  !--
Ax (Ey (R_ind(x, y) & F(y))  --> ^Ez (R_ind (x, z) & G(z))).
```

(Принцип уникальности ассоциированного объекта для любого индивидуализирующего отношения R_ind.

Примеры:
Человек имеет единственное место основной работы;
Юридическое лицо имеет единственный юридический адрес;
Страна имеет единственную столицу;
и т. п.)

Все схемы вывода реализуются процедурно и, следовательно, не требуют использования дополнительных изобразительных средств ни в языке индексирования, ни в тезаурусе.

3.3. Язык индексирования.

В качестве языка индексирования используется язык семантических графов (СГ). С точки зрения соотношения с выразительными средствами естественного языка можно различить два способа выражения смысла на языке СГ. В первом случае узлы СГ представляют только объекты, а дуги - все виды отношений между объектами, независимо от того, выражаются они в естественном языке лексическими или грамматическими средствами. (При этом, конечно, предполагается, что в конкретной предметной области различие между именами объектов и другими категориями терминов может быть проведено достаточно четко.) Во втором случае разграничение содержания узлов и дуг в большей степени ориентировано на естественный язык в том смысле, что узлы рассматриваются как основное средство представления лексических единиц естественного языка, а дуги - как средство представления, главным образом, грамматически выражаемых отношений. При этом используется весьма ограниченная номенклатура дуг с обобщенным значением, представляющих ту часть грамматических средств естественного языка, которая ответственна за оформление внутренней структуры именных групп и их связей с единицами глагольного ряда (аналог предложно-падежной системы).

Здесь используется язык индексирования, соответствующий СГ второго типа. Примеры формализации смысла текста, представленные в специальной демонстрационной нотации показаны на рис. 4.

Кроме перечисленных, при переходе от логического языка к языку ИЛП приходится решать ряд других вопросов принципиального характера: выбор уровня атомизации лексики (уровня семантических примитивов); целесообразность сохранения тех или иных поверхностно-семантических различий в лексике (пример - процессные термины); определение номенклатуры различаемых лексических отношений и семантических валентностей и др.

Text Tracing Analysis

Text

Вал вращается асинхронным двигателем с фазным ротором мощностью 400 вт

Text 1 Sentence 3 Pair

Abstract

```
ОБЪЕКТ  1:  ВАЛ
ОБЪЕКТ  2:  АСИНХРОННЫЙ_ДВИГАТЕЛЬ,  ( МОЩНОСТЬ  = 400 )
ОБЪЕКТ  3:  ФАЗНЫЙ_РОТОР

                    ВРАЩАТЬ
                            ЧТО: ОБЪЕКТ 1
                    ПОСРЕДСТВОМ: ОБЪЕКТ 2 .
          ОБЪЕКТ  2  ИМЕТЬ_ЧАСТЬЮ  ОБЪЕКТ  3
```

Рис. 4. Пример формализации содержания текста (использован демонстрационный язык).

4. МЕТОДЫ РЕАЛИЗАЦИИ ИНФОРМАЦИОННО-ЛИНГВИСТИЧЕСКОГО ПРОЦЕССОРА

4.1. Общая архитектура ИЛП ([23]).

На уровне реализации наибольшую сложность представили следующие аспекты:

- организация межуровневого взаимодействия при выделении уровней морфологического, синтаксического, локально-семантического и межфразового анализа как относительно самостоятельных функциональных компонент;
- обеспечение эффективности;
- структура тезауруса, методология его наполнения и программная реализация (организация наборов данных).

При определении общего подхода к реализации учитывались два принципиальных обстоятельства:

- недостаточная модельная разработанность процедур внутриуровневого анализа (особенно на семантических уровнях) и необходимость сосредоточить внимание на разработке этих процедур;
- ограниченность аппаратных ресурсов при принципиальной для данной работы ориентированности на серийную, массово доступную технику (класса AT 286).

В соответствии с этими посылками было принято решение не использовать в качестве базового механизма анализа механизм *backtracking'a*. После этого архитектура программного комплекса определилась как последовательность одноуровневых функционально замкнутых анализаторов, связанных только через данные при сохранении на выходе всех неоднозначностей анализа. Фильтрация неоднозначностей, как правило, осуществляется средствами следующего уровня. При неудаче разрешения неоднозначности она может быть передана дальше, лишь будучи переформулированной в терминах уровня текущей обработки. Отсюда очевидная многопроходность анализа.

Критическим параметром для системы в целом является скорость обработки. Для отдельных уровней - следующие аспекты обработки.

Уровень морфологического анализа - время доступа к словарю основ.

Уровень синтаксического анализа - предотвращение комбинаторного взрыва при переборе вариантов.

Уровень локально-семантического анализа и уровень межфразового анализа - время доступа к тезаурусу.

Для последних двух уровней существенно, что сами процедуры анализа представляют собой систему вложенных циклов, причем обращения к тезаурусу выполняются всегда из внутренних циклов. При этом обращение к тезаурусу, как правило, представляет собой не просто поиск нужной словарной статьи, а вызов машины прямого вывода, которая сама реализуется как система вложенных циклов обращения к тезаурусу. Это предопределяет крайне жесткие требования относительно допустимого времени доступа к словарной статье тезауруса. С учетом сказанного, единственным практически приемлемым решением оказывается размещение рабочих наборов тезауруса целиком в оперативной памяти. Это, в свою очередь, предопределяет крайне жесткие требования к размерам словарной статьи тезауруса и способам организации тезаурусных наборов.

4.2. Методы реализации модуля концептуального анализа ([23]).

Общая схема модуля показана на рис. 5.

Процедура концептуальной интерпретации синтаксического графа состоит в последовательном просмотре парных связей "хозяин-слуга" в направлении снизу вверх, слева направо. Интерпретатор представляет собой набор блоков-специалистов, каждый из которых ответственен за интерпретацию связей одного определенного типа. Тип связи, как правило, определяется парой <семантическая категория хозяина, семантическая категория слуги>. Соответствие между типом связи и номером интерпретирующего блока устанавливается с помощью управляющей таблицы. Назначение основных блоков:

1. Навешивание наименований признаков на объекты.
2. Навешивание наименований групп признаков и процессных признаков.
3. Анализ связей внутри объектных групп.
4. Установление связей типа "признак-значение".
5. Присоединение определителей.
6. Установление связей по валентности.
7. Модификация значений и определителей.
8. Исключение избыточных элементов.

Рис. 5. Общая схема модуля концептуальной интерпретации.

Схема программной реализации семантического интерпретатора показана на рис. 6. Интерпретатор состоит из трех блоков: блока управления, блока анализа и блока исполнения.

Блок управления находит очередной терминальный узел синтаксического графа, определяет по управляющей таблице нужного специалиста-анализатора и передает ему управление.

Блок анализа представляет собой набор процедур, специализированных по типам связи. Внутри некоторых блоков предусмотрена дальнейшая специализация в зависимости от семантического типа дескрипторов, связанных данной подчинитель-

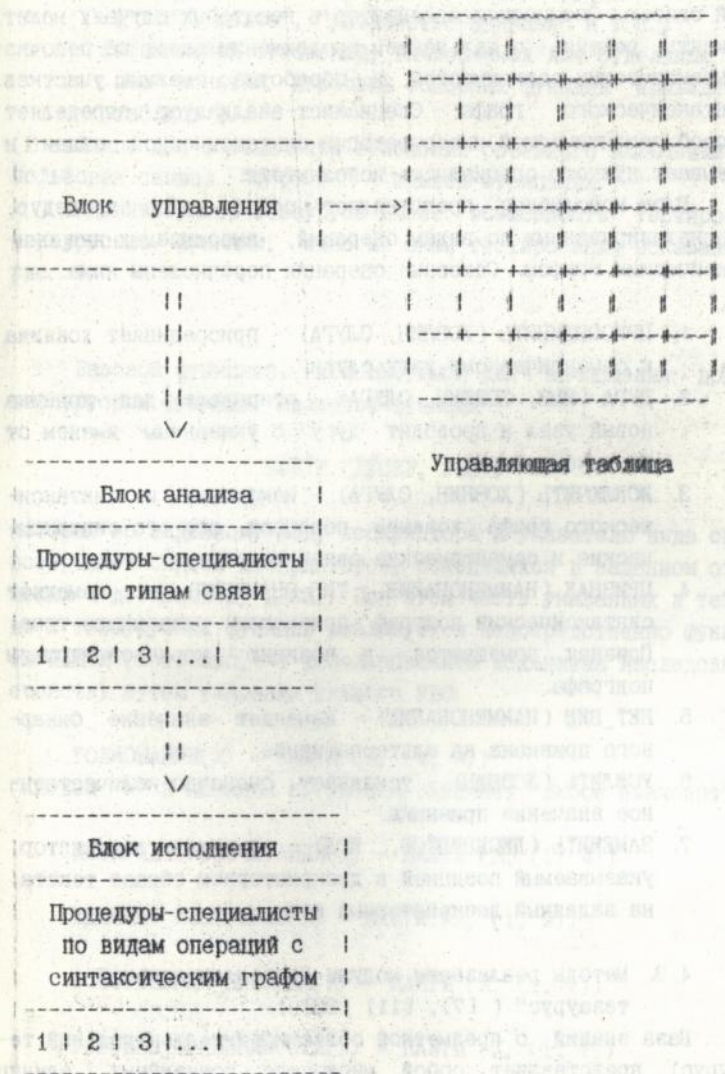

Рис. .6. Организация семантического интерпретатора.

ной связью. Специалист-анализатор в некоторых случаях может принять решение о дальнейшем продвижении вверх по цепочке синтаксических зависимостей и обработке смежных участков синтаксического графа. Специалист-анализатор определяет способ семантической интерпретации синтаксической связи и вызывает нужного специалиста-исполнителя.

Блок исполнения представляет собой набор процедур, специализированных по видам операций, выполняемых над синтаксическим графом. Основные операции перечислены ниже.

1. ПРИСОЕДИНИТЬ (ХОЗЯИН, СЛУГА) - присоединяет хозяина к семантическому узлу слуги.
2. ДУГА (ИМЯ, ХОЗЯИН, СЛУГА) - открывает для хозяина новый узел и проводит дугу с указанным именем от хозяина к слуге.
3. ИСКЛЮЧИТЬ (ХОЗЯИН, СЛУГА) - исключает из синтаксического графа хозяина, перенося все его синтаксические и семантические связи на слугу.
4. ПРИЗНАК (НАИМЕНОВАНИЕ, ТИП_ЗНАЧЕНИЯ) - заменяет синтаксический подграф признаком указанного типа. Признак помещается в позицию корневой вершины подграфа.
5. НЕТ_БИН (НАИМЕНОВАНИЕ) - изменяет значение бинарного признака на альтернативное.
6. УСИЛИТЬ (ХОЗЯИН) - усиливает оценочно-количественное значение признака.
7. ЗАМЕНИТЬ (ДЕСКРИПТОР, КОД) - заменяет дескриптор, указываемый позицией в дескрипторном образе текста, на заданный дескрипторный код.

4.3. Методы реализации модуля "интеллектуальный тезаурус" ([7], [11] [23]).

База знаний о предметной области (интеллектуальный тезаурус) представляет собой множество понятийных единиц ("дескрипторов") с заданным на них набором отношений.

Интерфейс собственно анализатора с базой знаний оформлен как библиотека функций, обеспечивающих вычисление

- полного набора объемных отношений между понятиями;

- фиксированного набора ассоциативных отношений ("часть-целое", "устройство-функция" и т. п.)
- специальных отношений, вычисляемых для отдельных типов понятий. Перечень основных функций приведен в таблице 2.

(В таблице для обрзначения отношения объемного включения использован символ "<"; "^" - символ отрицания.)

Администратор тезауруса имеет возможность тестировать тезаурусные функции; одно из окон тестирования показано на рис. 7.

Базовой функцией, используемой для вычисления других тезаурусных функций, является функция

НАЙТИ (ДЕСКР, УВС),

которая по заданному коду дескриптора и указателю вида связи возвращает список дескрипторов, находящихся в заданном отношении к дескриптору ДЕСКР. При этом часть указанных в таблице 2 тезаурусных функций реализуется непосредственно функцией НАЙТИ (возможно, с использованием механизма наследования свойств) путем указания нужного УВС:

ТОЛКОВАНИЕ(X) = НАЙТИ (X, (4, *))

(символ "*" в позиции аргумента означает "любое значение".)

ДЕСКРИПТОРОСОЧЕТАНИЯ(X) = НАЙТИ (X, (1, 4))

ЕДИНИЦА_ИЗМЕРЕНИЯ(X) = НАЙТИ (X, (1, 9))

СОБСТВЕННЫЙ_ПРИЗНАК(X) = НАЙТИ (X, (1, 1))

УСЛОВИЕ_ПРИМЕНИМОСТИ(X) = НАЙТИ (X, (1, 1))

ЧАСТИ(X) = НАЙТИ (X, (7, 1))

ЦЕЛОЕ(X) = НАЙТИ (X, (6, 1))

Рассмотрим подробнее способ реализации некоторых из более сложных функций.

Таблица 2

Имя функции	Область определения функции	Возвращаемое значение
ВЫШЕ_1(X)	X - любой дескриптор	Множество всех дескрипторов Y, таких что X < Y
ВЫШЕ_2(X1, X2)	X1, X2 - дескрипторы одной категории	ДА - если X2 < X1 НЕТ - если ^(X2 < X1)
СОВМЕСТИМЫ(X1, X2)	X1, X2 - имена объектов (базовые или или производные свойства)	ДА - если пересечение X1 и X2 не пусто НЕТ - если пусто
ТОЛКОВАНИЕ(X)	X - производное свойство	Запись, представляющая толкование дескриптора X
СОБСТВЕННЫЙ_ПРИЗНАК(X)	X -значение классификационного признака	Дескрипторный код наименования признака, значением которого является X
СПИСОК_ЗНАЧЕНИЙ(X)	X - наименование классификационного признака	Множество дескрипторов-значений признака X
УСЛОВИЕ_ПРИМЕНИМОСТИ(X)	X - наименование признака	Дескриптор - условие применимости признака X
ЕДИНИЦА_ИЗМЕРЕНИЯ(X)	X - наименование количественного признака	Дескрипторный код стандартной единицы измерения
ИЗМЕРЯЕМАЯ_ВЕЛИЧИНА(X)	X - единица измерения	Множество наименований количественных признаков, измеряемых единицей X

Таблица 2 (окончание)

Имя функции	Область определения функции	Возвращаемое значение
ЦЕЛОЕ(X)	X - имя объекта	Дескриптор, представляющий класс объектов, имеющих частью X
ЧАСТИ(X)	X - имя объекта	Множество всех дескрипторов, являющихся частью X
АССОЦИАЦИИ_ДЛЯ(X)	X - имя объекта или процесса	Множество всех дескрипторов, связанных с X каким-либо ассоциированным отношением
АССОЦИИРОВАННОЕ ОТНОШЕНИЕ_ДЛЯ(X1, X2)	X1, X2 - имена объектов или процессов	Дескриптор, представляющий ассоциированное с заданной парой отношение
ПРИЗНАКИ_ДЛЯ(X)	X - имя объекта	Множество всех наименований признаков, применимых к объектам класса X
МОДЕЛЬ_УПРАВЛЕНИЯ(X)	X - отношение или процесс, действие, характеризуемые моделью управления	Множество пар вида <v, u> (v - имя валентности, u - условие заполнения валентности)

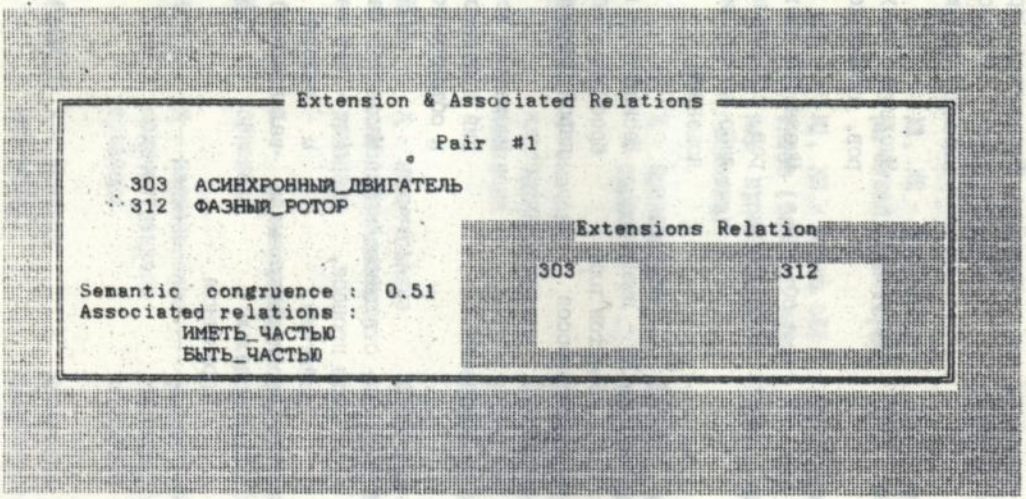

Рис. 7. Результаты вычисления базовых тезаурусных функций для пары понятий *асинхронный двигатель* и *фазный ротор*.

Функция ВЫШЕ_1(X) строит развертку дескриптора X; функция СОВМЕСТИМЫ (X1, X2) строит развертки для X1 и X2 (соответственно П1 и П2) и проверяет наличие в них несовместимой пары дескрипторов.

Вычисление функции ВЫШЕ_2(X1, X2) в реализации встроено в функцию СОВМЕСТИМЫ(X1, X2). Последняя возвращает пару (FI0, FI1), где FI0 указывает собственно результат вычисления совместимости а FI1 - сопутствующий параметр, семантика которого зависит от значения основного параметра FI0.

Правила построения развертки можно описать следующим образом:

```
1.      d ∈ П(d)
          3.1
2.1.    d     --->  УП(ПРИЗН(d))

          3.2
2.2.    d     --->  ДЕФ(d)

          3.3
2.3.    d     --->  УП(ПРИЗН(ДЕФ1(d)))

          3.4  3.5  3.6
2.4.    d  , d  , d     --->  УП(ДЕФ1(d))

          3.7
2.5.    d     --->  0

          5.1  5.2
2.6.    d  , d     --->  0

          3.*
2.7.    d     --->  ЗЗв1(d)
```

Имена функторов здесь имеют следующее значение:

ПРИЗН(X) - наименование признака, значением которого является X;

УП(X) - условие применимости для признака X;

ДЕФ(X) - список дескрипторов, входящих в толкование X;

ДЕФ1(X) - первый дескриптор толкования X;

ЭЗв1(X) - список дескрипторов, связанных с X отношением непосредственной эмпирической зависимости.

Правило вида d ---> F(d) означает, что при наличии в развертке П дескриптора d должна быть выполнена операция пополнения развертки

$$П \longrightarrow П \cup F(d).$$

Правило 2.3 сформулировано в предположении, что пополняющие дескрипторы для всех компонент ИЛИ-толкования одинаковы. Это ограничение вводится чтобы избежать рекурсии при вычислении развертки и с учетом того, что ИЛИ-толкование в реализации рассматривается как вспомогательное (а не регулярное) средство введения новых дескрипторов в тезаурус.

Алгоритм построения развертки может быть описан следующим образом.

1. Включить в развертку исходный дескриптор d.

2. Перебирая последовательно дескрипторы развертки, применять к каждому процедуру пополнения в соответствии с правилами, указанными для данного типа дескрипторов.

3. Если список дескрипторов исчерпан - конец.

Алгоритм, реализующий вычисление функции ПРИЗНАКИ_ДЛЯ(X), выполняет следующие действия.

1. Строит развертку дескриптора X.

2. Просматривает дескрипторы тезауруса в порядке кодов и выбирает наименования признаков (d), кроме фиктивных признаков и признаков, выделяющих заданный класс X.

3. Для каждого выбранного дескриптора d проверяет совместимость его условия применимости и дескриптора X.

4. Если условие п.4 выполнено - включает дескриптор в фрейм-список.

В тезаурусе дескрипторы распределены по семантическим категориям (СК), внутри категорий - по семантическим типам (СТ). Состав словарной статьи дескриптора переменный и зависит от семантической категории дескриптора, семантического типа и, возможно, также от значения более детальных характеристик.

Словарная статья тезауруса реализуется следующей схемой:

```
<словарная статья> ::= <описатель><список связей>

<описатель>        ::= СК СТ ДХ В1 В2 В3 В4

<список связей>    ::= <связь> | <список связей><связь> |
                       | <список связей><значение>

<связь> ::= <указатель вида связи> АССОЦИИРОВАННЫЙ_ДЕСКРИПТОР
<указатель вида связи> ::= УВС1 УВС2
<значение> ::= КОД | ВЕЩЕСТВЕННОЕ_ЧИСЛО |
             | ВЕЩЕСТВЕННОЕ_ЧИСЛО ВЕЩЕСТВЕННОЕ_ЧИСЛО |
             | ЦЕЛОЕ_ЧИСЛО ЦЕЛОЕ_ЧИСЛО ЦЕЛОЕ_ЧИСЛО |
             | СТРОКА_СИМВОЛОВ
```

Здесь: СК - семантическая категория дескриптора;
СТ - семантический тип дескриптора;
ДХ, В1, В2, В3, В4 - элементы описания дескриптора, семантика которых зависит, главным образом, от категории и типа;
УВС1, УВС2 - коды, определяющие вид связи (далее будем эту пару обозначать как УВС).

Список категорий и типов приведен в таблице 3. В категории 4 собраны служебные ("строевые") элементы понятийной системы. В категории 5 собраны дескрипторы, словарное описание которых может быть ограничено набором коротких элементов; соответственно процедура анализа на совместимость их с другими дескрипторами и между собой может быть предельно упрощена. Так, для терминов свободной сочетаемости (СКТ-5.1) указывается только семантическая категория и тип. В категории 5 представлены, как правило, прилагательные с достаточно широким и контекстно обусловленным значением: *автоматический, автономный, цифровой, активный, механический, статический, динамический, электрический, линейный* и т. п.

На рис. 8 приведен фрагмент тезауруса, в котором представлены словарные статьи дескрипторов (коды с 318 по 322).

Для определения структуры тезауруса можно использовать метод самоописания, представив систему словарных характеристик тезауруса в виде дерева словарных признаков. При таком подходе сам тезаурус рассматривается как особая предметная область, а все средства ведения тезауруса могут использоваться для определения и корректировки самой системы словарного описания. В принципе этот прием может быть распространен и на любые другие словари системы, что позволяет унифицировать средства работы со словарями.

Таблица 3

Семантическая классификация дескрипторов

Семантическая категория Семантический тип	Пример
1. Наименование признака	
1.1. Качественный	*цвет*
1.2. Бинарный	*истинность*
1.3. Целочисленный	*число полюсов*
1.4. Количественный	*вес*
1.5. Строковый	*марка*
1.6. Процессный	*частота*
1.7. Без значений	*структура*
2. Наименование группы признаков	*параметр*
3. Объектный дескриптор	
3.1. Базовые свойства	*газообразный*
3.2. Определяемый И-толкованием	*лед*
3.3. Определяемый ИЛИ-толкованием	
3.4. Определяемый числовым значением признака	*весом 1 т*
3.5. Определяемый строковым значением признака	*Москва*
3.6. Определяемый кодовым значением признака	*тяжелый*
3.7. Определяемый R-толкованием	*коллекторный*
4. Служебный терм	
4.1. Единицы измерения	
4.2. Оценочно-количественные значения признака	*большой*
4.3. Модификаторы оценочно-количественных значений	*очень*
4.4. Термы со значением отрицания	*не, без*
4.5. Кванторные и количественные определители	*многие*
4.6. Семантически избыточные элементы	*процесс*
4.7. Указатели роли	*с помощью*
4.8. Индикаторы референции	*этот*

Таблица 3 (окончание)

Семантическая категория Семантический тип	Пример
5. Дескриптор свободного употребления	
5.1. Дескрипторы свободной сочетаемости	*максимальный*
5.2. Кустовые	*статический - динамический*
5.3. Квазиместоимения	*что*
6. Процессный дескриптор	
6.1. Типа ГЛАГОЛ-СВЯЗКА	*был*
6.2. Процессный классификатор	
6.3. Типа ИЗМЕНЕНИЕ	*нагрев*
6.4. Типа ПРЕВРАЩЕНИЕ	*авария*
6.5. Значения процессных признаков	*быстрый*
7. Статическое отношение	
7.1. Предикаты кореференции	*представляет собой*
7.2. Объектные Т-отношения	*расстояние*
7.3. Объектные Ф-отношения	*быть частью*
7.4. Отношения величины	*превосходит*
7.5. Процессные отношения	*предшествует*

Простые термины - наименования признаков и базовые свойства - организуются в тезаурусе в дерево (лес) признаков при помощи ссылок ВВЕРХ - элементы описания "условие применимости" (УП) и "собственный признак" (ПРИЗН) для семантических категорий СК-1 и СК-3, соответственно. Базовые свойства дополнительно могут быть связаны ссылками, представляющими эмпирические зависимости. Они отличаются от ссылок, образующих дерево признаков тем, что соединяют однотипные вершины, тогда как ссылки УП и ПРИЗН соединяют разнотипные вершины - наименования признаков и базовые свойства.

```
Environment   Analyse   Thesaurus   Vocabulary   I/O Edit   UserThesaur   Quit

 318    3 2 0000 0      ТЯГОВЫЙ ДВИГАТЕЛЬ
              4   0    301  ЭЛЕКТРОДВИГАТЕЛЬ
              4   0    317  МОЩНОСТЬ СВЫШЕ 100 КВТ

 319    3 4 0000 0      МОЩНОСТЬ ОТ 1000 ДО 10000 ВТ
              4   0    304  МОЩНОСТЬ
              4   2     1.00000E+03
              4   2     1.00000E+04

 320    1 1 0000 0      МАРКА
              1   1    143  ИЗДЕЛИЕ
              1   4    321  МАРКА=ЭР-57
              1   3    376  НАДЕЖНОСТЬ

 321    3 5 0000 0      МАРКА=ЭР-57
              4   0    320  МАРКА
              4   5    401
              1   4    322  ЭЛ.МАШИНА МАРКИ ЭР-57

 322    3 2 0000 0      ЭЛ.МАШИНА МАРКИ ЭР-57
              4   0    275  ЭЛЕКТРИЧЕСКАЯ_МАШИНА
              4   0    321  МАРКА=ЭР-57
Line  1371 Alt-C-Find code|Alt-W/N/P-Find word/next/prev|Esc-exit Scroll UP&Down
```

Рис. 8. Фрагмент тезауруса (Предметная область "Электрические машины")

В результате получается базовая структура, которая может быть представлена в виде ациклического орграфа; в терминах этой структуры определяется основное понятие развертки.

> Развертка дескриптора d определяется как множество простых следствий d, определяемых прямыми связями между дескрипторами
>
> $$П(d) = \{d' | d \dashrightarrow d'\}.$$

Для установления несовместимости двух дескрипторов d1 и d2 необходимо и достаточно, сравнив их развертки П(d1) и П(d2), обнаружить в них пару дескрипторов, представляющих разные значения одного и того же признака.

Для обеспечения навигационных операций по дереву признаков предусмотрены также ссылки ВНИЗ - ВЛЕВО - к первому слева подчиненному дескриптору и ВПРАВО - к ближайшему соседу справа. Эти ссылки при вводе дескрипторов устанавливаются автоматически.

Предусмотренные в тезаурусе типы формальных толкований указаны в табл. 3. Используемый тип толкования определяется семантическим типом дескриптора (внутри категории СК=3). Относительно структуры толкований подразумеваются следующие прагматически оправданные ограничения.

1. Однородность толкования. В каждом толковании может быть реализован лишь один вид отношений между элементами толкования:

- для И-толкования - конъюнкция;
- для ИЛИ-толкования - дизъюнкция;
- для толкований через значение признака - связь между наименованием и значением признака;
- для R-толкования - связь между именем отношения и объектом отношения;

В случае необходимости построить неоднородное толкование оно строится в несколько последовательных шагов. Так, если требуется построить толкование со связками И и ИЛИ одновременно, например, вида

$$D \text{ -- } (D_1 \vee D_2) \& D_3 ,$$

то сначала определяется дескриптор

$$D' \text{ -- } D_1 \vee D_2$$

(строится ИЛИ-толкование).

Затем дескриптор D определяется как

$$D' \& D_3$$

(строится И-толкование).

Аналогично и в других подобных случаях.

2. В R-толкованиях могут использоваться только термины двухместных отношений.

3. В ИЛИ-толкованиях не могут использоваться дескрипторы, вводимые И-толкованиями.

Логический смысл этого ограничения состоит, очевидно, в том, что любое толкование, использующее связки И и ИЛИ должно быть приведено к конъюнктивной нормальной форме.

ЗАКЛЮЧЕНИЕ

Накопленный нами и рядом других исследовательских групп опыт иссследований и разработок, а также быстрый прогресс в технологии программирования и микроэлектронике создали, по нашему мнению, достаточные предпосылки для организации мощного прорыва в области систем обработки языковых данных. Для задачи собственно автоматического анализа текста здесь остается еще, во-первых, весьма объемная работа по совершенствованию методов анализа (но уже, можно надеяться, в русле сформировавшегося круга идей и методов) и, во-вторых, также весьма трудоемкая работа по созданию и ведению комплекса словарной поддержки - объемом не менее 100 тыс. единиц как для грамматических, так и для понятийных словарей. Полагаем, что на этой основе можно говорить о предстоящей в ближайшее десятилетие интеграции в едином программно-технологическом комплексе средств анализа и синтеза не только текста, но и устной речи. Как результат - можно ожидать вовлечения в компьютерную технологию совершенно новых и, возможно, на первый взгляд неожиданных сфер человеческой деятельности. И прежде всего от степени осознания этой перспективы членами профессионального сообщества зависит, когда и как будет совершен этот прорыв.

Основные публикации по теме диссертации

1. В. Ш. Рубашкин. Роль математики в развитии понятийного аппарата науки // Математизация знания. - М.: Институт математики СО АН СССР. - 1968. - С. 110 - 123.
2. В. Ш. Рубашкин. Познание и язык // Вопросы философии. - 1970. - N 9. - С. 50-59.
3. В. Ш. Рубашкин. Математическая логика и язык науки // Вопросы философии. - 1973. - N 1. - С. 112-122.
4. В. Ш. Рубашкин. О грамматических средствах информационных языков // Информационные языки. - М.: - 1975. - С. 191-222.
5. В. Ш. Рубашкин. Признак и значение // Научно-техническая информация. Сер. 2. - 1976. - N 3. - С. 3-10.

6. В. Ш. Рубашкин. О семантике определительных конструкций // Лингвистические проблемы функционального моделирования речевой деятельности. Вып. 4. - Л.: ЛГУ, 1979. - С. 46-55.
7. В. Ш. Рубашкин. О семантической сочетаемости лексики // Научно-техническая информация. Сер. 2. - 1981. - N 2. - С. 21-29.
8. И. С. Добронравов, Д. Г. Лахути, В. Ш. Рубашкин, Н. Г. Сорокина. Синтаксис как средство смыслоразличения в документальных ИПС с автоматическим индексированием // Научно-техническая информация. Сер. 2. - 1981. - N 3. - С. 17-24.
9. В. Ш. Рубашкин. Об одном типе правил смыслоотождествления // Лингвистические проблемы функционального моделирования речевой деятельности. Вып. 5. - Л. : ЛГУ, 1982. - С. 143-150.
10. В. Ш. Рубашкин. О точности языка классификационных систем // Теория классификации и анализ данных: Материалы Всесоюзного совещания. - Новосибирск - ВЦ СО АН СССР, 1982. - С. 20-28.
11. В. Ш. Рубашкин. Формирование массивов данных путем диалога с системой классификации // Вопросы информационной теории и практики. - Вып. 47. - М.: ВИНИТИ. - 1982. - С. 82-99.
12. В. Ш. Рубашкин, Н. Г. Сорокина. Что нужно для распознавания парных связей? // Структурная и прикладная лингвистика. - Вып. 2. - Л.: ЛГУ, 1983. - С. 162-169.
13. В. Ш. Рубашкин. Об одной закономерности выражения денотативного тождества имен в связном тексте // Семиотические аспекты формализации интеллектуальной деятельности: Тезисы докладов и сообщений школы-семинара "Телави-83" - М. - 1983. - С. 233-235.
14. В. Ш. Рубашкин. Модели семантической сочетаемости лексики как инструмент анализа связного текста // Международный семинар по машинному переводу : Тезисы докладов. - М. - 1983. - С. 192-193.
15. В. Ш. Рубашкин. О методах анализа связного текста // Вопросы информационной теории и практики. - Вып. 49. - М.: ВИНИТИ. - 1983. - С. 58-73.

16. В. Ш. Рубашкин. Логический язык для описания сочетаемости терминов-имен свойств // Научно-техническая информация. Сер. 2. - 1984. - N 4. - С. 12-17.
17. В. Ш. Рубашкин. Логические модели терминосистем // Научно-техническая информация. - Сер. 2. - 1984. - N 12. - С. 10 - 14.
18. В. Ш. Рубашкин. Основные черты интеллектуальных информационных систем // Проблемы развития и освоения интеллектуальных информационных систем. (Тезисы докл. и сообщ. к Всес. конф. 11-13 ноября 1986 г.) Секция 1. Методология освоения и развития интеллектуальных систем. - Новосибирск, 1986. - С. 49 -50.
19. В. Ш. Рубашкин. Некоторые проблемы формализации знаний в интеллектуальных системах // Логика и системные методы анализа научного знания. Тезисы докл. к IX Всес. совещанию по логике, методологии и философии науки. Харьков, 8 - 10. X. 1986. Секция 1 - 5. - М., 1986. - С. 281 -282.
20. В. Ш. Рубашкин. Интеллектуальный тезаурус как модель терминологической системы // Вторая Всес. конф. по созданию машинного фонда русского языка (Тезисы докл.) - М., 1987. - С. 151 - 152.
21. Д. Г. Лахути, В. Ш. Рубашкин. Средства и процедуры концептуальной интерпретации входных сообщений на естественном языке // Известия АН СССР. Техническая кибернетика, 1987, N 2. - С. 49 - 59.
22. В. Ш. Рубашкин. От логического языка к языку интеллектуальной информационной системы // Семиотические аспекты формализации интеллектуальной деятельности. Всес. школа-семинар. Боржоми, 22 - 30 апреля 1988. Тезисы докл. и сообщ. - М., 1988. - С. 156 - 159.
23. В. Ш. Рубашкин. Представление и анализ смысла в интеллектуальных информационных системах. - М.: Наука, 1989. - 190 стр.
24. В. Ш. Рубашкин. Информационно-лингвистические процессоры: возможности реализации и анализ применения // Компьютер в музее, музей в компьютере. - Москва-Милан - 1991. - С. 105 -110.

Бесплатно

СПбГИК. Зак. 295.
Тир. 100. 4.8.92.