

0-741755

На правах рукописи

СУЛЕЙМАНОВ Джавдет Шевкетович

**СИСТЕМЫ И ИНФОРМАЦИОННЫЕ
ТЕХНОЛОГИИ ОБРАБОТКИ ЕСТЕСТВЕННО-
ЯЗЫКОВЫХ ТЕКСТОВ НА ОСНОВЕ
ПРАГМАТИЧЕСКИ-ОРИЕНТИРОВАННЫХ
ЛИНГВИСТИЧЕСКИХ МОДЕЛЕЙ**

05.13.14 – системы обработки информации и управления

А в т о р е ф е р а т
диссертации на соискание ученой степени
доктора технических наук



КАЗАНЬ — 2000

Работа выполнена на кафедре теоретической кибернетики Казанского государственного университета

Официальные оппоненты:

доктор физико-математических наук,
профессор

М.Г.Мальковский

доктор технических наук,
профессор

В.Ф.Хорошевский

доктор технических наук,
профессор

А.П.Тунаков

Ведущая организация: Институт высокопроизводительных вычислений и баз данных Миннауки РФ

Адрес: С.-Петербург, Фонтанка 118.

Защита состоится " 10 " марта 2000 г. в 10 часов на заседании диссертационного совета ССД 063.09.02 Казанского государственного технического университета им. А.Н. Туполева по адресу: г.Казань, ул. К.Маркса, 10, КГТУ.

С диссертацией можно ознакомиться в библиотеке университета

Автореферат разослан " _____ " _____

— НАУЧНАЯ БИБЛИОТЕКА КГУ



0000011105

Ученый секретарь
Совета

Р.Т. Сиразетдинов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. Экспоненциальный рост объема информации, накапливаемой в различных базах знаний, а также неформатированных машиночитаемых текстов, с одной стороны, и отсутствие удовлетворительных средств их обработки - с другой, ставят задачу построения средств автоматизации обработки информации на основе принципиально новых подходов и технологий. В этой связи актуальной является разработка лингвистических моделей, достигающих эффективности за счет их ориентированности на определенный класс лингвистических задач. Лингвистические модели - это, по сути, научно-прикладная область фундаментальных исследований, помогающая строить системы, которые обрабатывают: анализируют, генерируют, интерпретируют и трансформируют естественно-языковые (ЕЯ) тексты. Построение систем и информационных технологий на основе универсальной лингвистической модели является малоперспективным, с одной стороны, ввиду отсутствия универсальной формальной базы, и даже приемлемой полной формальной модели какого-либо языка, с другой стороны, из-за того, что реализация даже некоей модели, приближенной к универсальной, с применением современных технологий, будет неэффективной и малоприменимой по временным и емкостным характеристикам.

Известные лингвистические процессоры (ЛП), или системы обработки ЕЯ-текстов, как правило, созданы на основе лингвистической модели, включающей полный или ограниченный анализ морфологии, синтаксиса и семантики без учета специфики классов. Эффективность таких систем достигается за счет существенных ограничений либо на язык, либо на проблемную область, либо на обе эти составляющие. Системы и технологии обработки информации, рассматриваемые в данной работе, основаны на прагматически-ориентированных моделях, или моделях от цели, позволяющих перейти от сложных многоуровневых моделей с множеством связей к целевым моделям, включающим минимальный набор средств, разработанных исходя из принципа достаточности для решения задач.

Исследованные диссертантом прагматически-ориентированные лингвистические модели: (а) модель специальных грамматик как база проблемно-ориентированного семантического интерпретатора естественно-языковых текстов в управляемом контексте, (б) двухуровневая и генеративная модели морфологии татарского языка как основа двухуровневого морфологического анализатора и автоматизированного корректора текстов, (в) модель морфем как концептуально-формальная база для построения лингвопроцессоров и проведения научно-прикладных исследований — построены именно исходя из специфики задач соответствующей проблемной области, что и позволило разработать информационные технологии и реализовать лингвистические системы, активно и эффективно используемые на практике.

Целью диссертационной работы является разработка базовых принципов, архитектуры, технологии построения прагматически-ориентированных лингвистических моделей и реализации систем и информационных технологий обработки ЕЯ-текстов на их основе.

Диссертантом разработана методология и предложены методы и базовые принципы построения систем обработки естественно-языковой текстовой информации, основанных на прагматически-ориентированных моделях, или моделях "от цели", в отличие от универсальных "избыточных" моделей. Такой подход позволил минимизировать программно-информационное и лингвистическое обеспечение моделей обработки ЕЯ-текстов при построении лингвопроцессоров.

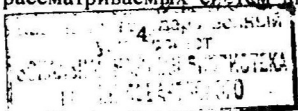
Для достижения цели проведен анализ существующих лингвистических моделей, систем и технологий автоматизированной обработки ЕЯ-текстов и решены следующие задачи:

- исследованы современные подходы и методы построения лингвистических моделей, а также определены принципы и критерии их построения,
- исследованы грамматические формализмы как основа лингвопроцессоров и системы обработки ЕЯ-текстов с точки зрения их результативности и эффективности,
- сформированы теоретические положения прагматически-ориентированного подхода и понятийный аппарат, необходимые для описания прагматически-ориентированных лингвистических моделей,
- разработаны прагматические модели естественного языка (русского, татарского) и алгоритмы анализа текстов,
- исследованы технологии создания и применения ЕЯ-систем и разработаны конкретные информационные технологии обработки естественно-языковых текстов,
- определены рациональные области применения ЕЯ-систем и созданы конкретные системы обработки информации различного назначения.

Полученные результаты зафиксированы в виде пакетов прикладных программ, базирующихся на алгоритмах и соответствующих программных средствах, а также в виде формальных моделей и соответствующих информационных ресурсов.

Научная новизна. В процессе исследований получены следующие научные результаты, выносимые на защиту.

Разработана методология и предложены методы и базовые принципы построения систем и технологий обработки текстовой информации, основанных на прагматически-ориентированных моделях, или моделях от цели, в отличие от универсальных "избыточных" моделей. Такой подход позволил минимизировать программно-информационное и лингвистическое обеспечение моделей обработки ЕЯ-текстов при построении лингвопроцессоров. Признание необходимости дифференцированного целевого подхода к построению лингвистических моделей предопределило одно из главных свойств рассматриваемых систем анализа ЕЯ-текстов - их



специализацию внутри определенного класса задач и, одновременно, адаптивность к задачам внутри данного класса и, соответственно, новизну частных решений.

Впервые поставлена и решена диссертантом задача *построения компьютерной модели морфологии татарского языка*. Новыми являются как сам прагматически-ориентированный подход к созданию лингвистических моделей, так и конкретные модели, построенные на основе разработанного подхода.

Новизна лингвистической модели как основы вопросно-ответных систем заключается в том, что в работе сформулированы и реализованы следующие базовые принципы ее построения в предположении, что система функционирует в режиме ожидаемого смысла, т.е. ожидаемых значений вопроса.

Принцип 1. Выделение смыслообразующих единиц в вопросно-ответном тексте, названных автором концептулами, с целью трансформации проблемы семантического анализа вопросно-ответного текста в проблему грамматического разбора в условиях использования детерминирующей роли контекста.

Принцип 2. Семантическая классификация вопросно-ответных текстов на основе типовых отношений: выделение типов вопросов и классов значений вопросов (т.е. ответов) для реализации детерминирующей роли контекста.

Принцип 3. Разработка индивидуальных концептуальных грамматик (ИКГ) семантических классов, отражающих смысловые конструкции ответов соответствующих классов и в совокупности составляющих концептуальную грамматику (КГ), служащую формальной основой для построения семантического интерпретатора, ориентированного на "слушающего".

Принцип 4. Сегментация вопросно-ответных текстов по минимальным смысловым конструкциям для рекурсивного применения правил концептуальной грамматики.

Новизна концептуально-формальных моделей обработки ЕЯ-текстов, раскрываемых в работе, прежде всего, определяется тем, что они разработаны и реализованы впервые для татарского языка. Новыми по сути являются все составляющие двухуровневого морфологического анализатора: база фонологических правил, база морфотактических правил, а также сам подход - не на генеративной, а на двухуровневой основе: лексической и поверхностной, что позволяет использовать анализатор одновременно как для морфологического анализа, так и генерации татарских словоформ. Морфологический корректор татарских текстов является первой системой, позволяющей осуществлять морфологическую коррекцию татарских текстов на основе орфографического словаря, словаря аффиксов и генеративных правил, обеспеченной удобным русско-татарским интерфейсом и системой подсказки возможных правильных вариантов словоформ. Прагматически-ориентированный подход позволил применить к построению формальной базы татарского морфологического корректора не декларативно-парадигматический подход, как это принято в флективных языках типа русского или английского,

а процедурально-генеративный подход, являющийся более эффективным для татарского языка с учетом явлений регулярности, рекурсии и потенциальной неограниченности длины словоформы в татарской морфологии.

Структурно-функциональная модель морфем, раскрываемая в диссертационной работе, не имеет аналогов и, в силу ее открытости, является универсальной моделью для "инвентаризации" морфем любого естественного языка. Впервые разработана формальная модель для описания значений аффиксальных морфем. Такие структурно-функциональные модели являются средствами, позволяющими практически осуществить переход от алгоритмического подхода к модельному, децентрализованному и асинхронному, подходу обработки естественно-языковых текстов.

Практическая ценность полученных результатов состоит в разработке технологии создания прагматически-ориентированных лингвистических моделей и реализации на основе предложенных моделей пакетов прикладных программ, активно используемых в информационных технологиях. Исследования, начатые автором в 1978 году в рамках Научного направления Казанского государственного университета "Математическое и программное обеспечение автоматизации интеллектуальной деятельности", в 1980 году были продолжены в соответствии с программой ГКНТ 0.80.10 и программой Минвуза СССР "Расширение и повышение эффективности применения ЭВМ в учебном процессе и научных исследованиях". В 1990 году работы осуществлялись совместно с МЭИ и Институтом микроэлектроники (г. Таганрог) по госбюджетной теме "Нейро-компьютеры", где автор являлся ответственным исполнителем одной из подтем: "Лингвистический процессор". Основные результаты диссертации получены в рамках НИР кафедры теоретической кибернетики Казанского государственного университета: "Математическое и алгоритмическое обеспечение автоматизации интеллектуальной деятельности", научного направления: "Концептуальные и алгоритмические аспекты человеко-машинного интерфейса" и Программы Академии наук Татарстана до 2005 года "Компьютерное обеспечение функционирования татарского языка как государственного. Концептуально-алгоритмическая модель".

Двухуровневый морфологический анализатор с 1996 года выполнялся на основе совместного международного проекта "Двухуровневый анализатор татарского языка" с Белкентским университетом (Турция, Анкара) при финансовой поддержке Европейской Программы "Наука за стабильность".

Практические разработки и реализация результатов диссертации с 1993 года осуществлялись по заказу Кабинета Министров РТ, а также в рамках Государственной программы Республики Татарстан по сохранению, изучению и развитию языков народов Республики Татарстан. Диссертант является руководителем работ по этим направлениям.

Разработанные методы, алгоритмы и пакеты программ активно используются в учебном процессе в Казанском государственном университете для обучения татарскому языку, в учебных курсах "Компьютерная лингвистика", "Человеко-машинный интерфейс", в научных исследованиях,

проводимых учеными Института языка, литературы и искусства АНТ, а также в редакциях татарских журналов и газет. Татарский морфологический корректор включен в состав системы автоматизированного распознавания текстов FineReader TATAR (фирма ABBYY, г.Москва), электронного русско-татарско-английского словаря (фирма МедиаЛингва, г.Москва), текстового редактора ТАТЕДКОР (СНИЛ АНТ и КГУ, г.Казань), программы проверки орфографии татарских текстов в среде Windows (СНИЛ АНТ и КГУ, г.Казань), широко используемых на практике.

Документы, подтверждающие внедрение и практическое использование результатов диссертации, прилагаются.

Личный вклад автора. Постановка решаемых задач, а также основные теоретические, экспериментальные и прикладные результаты принадлежат лично автору.

Методологический принцип "ожидаемости" в диалоговых системах был сформулирован и уточнялся в совместных обсуждениях с проф. Р.Г.Бухараевым.

Создание программно-информационного обеспечения, разработка семантических сценариев для описания значений морфем, а также реализация модели татарских аффиксальных морфем осуществлялось совместно с А.Р.Гатиатуллиным. Файлы фонологических и морфотактических правил татарской двухуровневой морфологии разработаны и реализованы совместно с Р.А.Гильмуллиним. Фрагменты концептуальной грамматики и интегрированная вопросно-ответная система разработаны и реализованы совместно с М.М.Аюповым. Соответствующие результаты являются частью кандидатских диссертаций, выполняемых под руководством автора и подготовленных или готовящихся к защите.

Автор искренне благодарен Р.Г.Бухараеву, А.И.Еникееву и Ф.С.Сафиуллиной, а также всем сотрудникам Совместной научно-исследовательской лаборатории АНТ и КГУ за оказанную помощь и поддержку.

Апробация работы. Результаты работы докладывались на 6 международных конференциях и семинарах за рубежом: на научном семинаре по интеллектуальным системам в Ростокском университете им. В.Пика (ГДР, 1990), на Международной конференции LP'96 по типологии языков (Чехия, г.Прага, 1996), на 9-ой Европейской конференции ML'9 по Обучению Машин (Чехия, г.Прага, 1997), на 6-ой Международной конференции UM'97 "Моделирование пользователя" (Италия, Киа Лагуна, 1997), на Международном семинаре "Формальная элегантность и естественная сложность морфологии" (Франция, г.Екс-ен-Провенс, 1997), на научном семинаре по ЕЯ-процессорам в Белкентском университете (Турция, г.Анкара, 1997), на Первом международном семинаре "Компьютеризация естественных языков" (Болгария, г.Варна, 1999); на 14 международных конференциях и семинарах, проводившихся в России и странах СНГ: на Всесоюзных конференциях по искусственному интеллекту с Международным участием (г. Минск, 1986, г.Тверь, 1992, г.Казань, 1996, г.Пушино, 1998), на

Международном рабочем семинаре в рамках КИИ-96 "Модели национальных языков" и "Искусственный интеллект в образовании" (г.Казань, 1996), на международных семинарах ДИАЛОГ: компьютерная лингвистика и ее приложения (г.Казань, 1995, г.Пушино, 1996, Ясная Поляна, 1997, Таруса, 1998, 1999) на Международной тюркологической конференции (г.Казань, 1992), на Международной конференции "Проблемы разработки языка Ортатюрк" (г.Ташкент, 1993), на Международной конференции "Компьютерный фонд тюркских терминов" (г.Шымкент, 1995), на Международной конференции "Языковая семантика" (г.Казань, 1997); а также на Всесоюзном совещании-семинаре по АОС "Проблемы автоматизации обучения персонала систем организационного управления" (г.Казань, 1983), на семинаре по диалоговым и обучающим системам в ИК им. В.М. Глушкова АН УССР (рук. проф. Довгялло А.М.) (г.Киев, 1984-85), на IY Всесоюзной конференции Диалог "Человек-ЭВМ" (г.Киев, 1985), на межвузовском научно-методическом семинаре "Математические модели и вычислительная техника в управлении учебным процессом высшей школы" (г.Рига, 1986), на национальной конференции "Искусственный интеллект в XXI веке" (г.Калининград, 1995), на различных республиканских, городских научных семинарах, итоговых научных конференциях КГУ (1988-1999), на телеконференции "Информационные технологии в гуманитарных науках" (КГУ, 1998). В 1987 году за цикл работ "Лингвистический процессор вопросно-ответных текстов" автор удостоен звания лауреата Республиканского конкурса научно-технического творчества молодежи.

Публикации. По теме диссертации опубликовано в открытой печати 73 работы (в том числе одна монография) суммарным объемом около 60 усл. печ. листов, отражающих основные научные результаты диссертации. Прикладные результаты представлены также в ряде научных отчетов по НИР, выполнявшимся под руководством диссертанта.

Структура и объем диссертации. Диссертация состоит из введения, пяти глав, заключения, списка литературы (включающего 213 названий) и приложений, включающих акты внедрения и использования разработок, демонстрационный материал и описания программных реализаций. Основной текст (без приложений) занимает 334 страницы.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы, сформулирована цель работы и определен перечень решаемых задач, указана их новизна, отмечены особенности подхода, раскрываемого в диссертационной работе, теоретическая и практическая ценность полученных решений и разработок, приведены основные защищаемые положения, коротко изложены основные положения диссертации.

Несмотря на обилие научной и технической литературы, посвященной описанию лингвистических процессоров, от узкоспециализированных до универсальных, ни одна из приведенных лингвистических задач в настоящее

время не имеет удовлетворительного решения, ни одна из программ человеко-машинного интерфейса, машинного перевода или запроса к базе данных не может претендовать на полное и окончательное решение этих проблем. Попытка найти решение через создание новых формализмов, ориентированных на описание лингвистических феноменов (морфологии, синтаксиса, семантики и др.), с годами привела к большому многообразию лингвистических моделей, различающихся только набором средств описания лингвистических феноменов, но не результативностью.

Решение проблемы построения *естественно-языковых лингвистических моделей* и создания на их основе *эффективных систем и информационных технологий обработки ЕЯ-текстов*, как это доказывается в диссертационной работе на примере разработки конкретных моделей, лежит не столько в области создания полных описаний ЕЯ (даже если это было бы возможно гипотетически), сколько в области концептуального осмысления подхода к построению лингвистической модели как к неотъемлемой части системы, составляющей единое целое вместе со всеми участниками обработки текста. Таким подходом, как мы считаем, является четкое базирование модели на прагматике системы, объединяющей всех участников вокруг целевой обработки ЕЯ-текста. Под участниками обработки текстов имеются в виду все привлекаемые ресурсы и субъекты, включая обрабатываемый текст на ЕЯ, пользователя, нелингвистические блоки системы, проблемную область, контекст и т.п.

Проведена следующая классификация *лингвистических моделей* по их прагматическим признакам, т.е. по цели разработки и сфере применения:

1) *когнитивные (семантико-контекстные) модели*, обеспечивающие глубинное проникновение в текущий контекст и трансформацию его с сохранением смысла как внутри одной модели, так и между разными моделями (например, системы машинного перевода, системы извлечения знаний);

2) *диалоговые запросно-ответные или интерактивные модели*, обеспечивающие естественно-языковой диалог автоматизированной системы с пользователем при запросах к системе или ответах пользователя на вопросы системы. Функция таких моделей, помимо анализа или синтеза ЕЯ-текстов, заключается в формировании корректного образа ответа на запрос и/или реакции на ответ пользователя;

3) *концептуально-формальные модели*, обеспечивающие целевую обработку текстов согласно соответствующих формальных правил определенного языкового уровня (собственно грамматики ЕЯ);

4) *концептуально-функциональные модели*, являющиеся наиболее полными описаниями определенного ЕЯ-уровня или уровней, обеспечивающими разработчиков структурно-функциональной, а также справочной информацией, необходимой при построении лингвопроцессоров.

Лингвистические модели класса 4 являются метамоделями по отношению к моделям других классов в том смысле, что они содержат "строительный материал", т.е. концептуальное и функциональное описание

единиц языковых уровней и их грамматик, из которых строятся модели трех первых классов.

Известные лингвистические процессоры, как правило, построены на основе общей лингвистической модели, включающей полный или ограниченный анализ морфологии, синтаксиса и семантики, но без первичного должного учета прагматической специфики классов задач, на решение которых они ориентированы. Эффективность таких систем достигается за счет существенных ограничений либо на язык, либо на проблемную область, либо на оба эти составляющие. Их можно назвать "усеченными моделями". Системы и технологии обработки информации, рассматриваемые в данной работе, основаны на прагматически-ориентированных моделях.

Прагматически-ориентированные модели, или модели от цели - это такие описания языка и процесса обработки естественно-языковых текстов, которые, в отличие от универсальных многоуровневых моделей с множеством связей, разработанных по принципу необходимости, включают минимальный набор средств, разработанный исходя из принципа достаточности для решения определенного круга задач. Такие модели строятся не за счет усечения тех или иных элементов языковых уровней универсальной лингвистической модели, а за счет установления целевых характеристик, изначально учитывающих достаточный набор языковых средств и детальность их описания исходя из методологических принципов "ожидаемости" и "детерминированности контекста". Принцип "ожидаемости" в общем случае является основой выбора и предопределения инструментария (формальной базы и словарей) для обработки входного текста, в то время как принцип "детерминированности контекста" естественным образом "диктует" параметры входного текста - его содержание, форму, лексикон.

Таким образом, *прагматический подход к построению лингвистических моделей* это, прежде всего, *концептуально-инструментальная технология, которая позволяет, с одной стороны, осуществлять адекватный подбор средств эффективной обработки ЕЯ-текста, с другой - детерминировать контекст и направлять формирование ожидаемого ЕЯ-текста.* Прагматика здесь проявляется и учитывается уже на уровне методологии, на уровне проектирования лингвистических моделей, а не только на уровне реализации, что, как правило, присуще проанализированным нами подходам к разработке систем обработки информации.

Прагматически-ориентированный подход устанавливает следующую технологию подбора необходимого "инструментария" (как процедурального, так и декларативного) для построения лингвистического процессора:

а) прежде всего, определяется класс моделей, следовательно, максимальный набор лингвистических и внелингвистических средств, который необходим для решения указанной задачи в рамках моделей данного класса (классы моделей 1-4),

б) определяется режим взаимодействия пользователя с системой (пакетный, интерактивный, активна система - пассивен пользователь, пассивна

система - активен пользователь и др.), следовательно, необходимый набор средств, определяющий схему диалога и поддерживающий данный режим,

с) определяется тип текста (текст для перевода с языка на язык, для перефразирования, для извлечения знаний, вопрос-ответ, запрос-ответ, для морфологического анализа, морфологической коррекции и др.), следовательно, определяется минимальная часть средств, выбранных пунктом (а).

В первой главе осуществляется аналитический обзор отечественных и зарубежных работ в области обработки естественного языка в аспекте прагматически-ориентированного подхода. Исследуются методы и подходы, близкие к прагматически-ориентированному подходу к разработке лингвистических моделей. Традиционный подход к анализу формальной базы систем обработки ЕЯ-текстов раскрывается в разделе 1.1.1. В разделе 1.1.2 проанализированы идеи и подходы, близкие к прагматически-ориентированному подходу построения лингвистических моделей. Анализу средств обработки ЕЯ-текстов в диалоговых системах посвящен раздел 1.2, включающий достаточно подробное рассмотрение особенностей диалоговых моделей в аспекте прагматически-ориентированного подхода. Здесь производится анализ особенностей двух режимов диалога системы с человеком: когда активен пользователь - пассивна система (запросно-ответные системы) и когда активна система - пассивен пользователь (вопросно-ответные системы). В современных исследованиях имеет место определенная поляризация: с одной стороны, разработаны очень простые грамматические модели, т.е. различного рода грамматики конечного состояния, которые поддерживают высокую эффективность обработки. С другой стороны, на противоположном конце шкалы находится целый ряд представлений мощных и лингвистически сложных формализмов, затрагивающих семантику, значения элементов языковых уровней. Анализ таких формализмов приводится в разделе 1.2.2. Анализу систем обработки ЕЯ-текстов на основе концептуально-формальной модели, а именно автоматизированных корректоров ЕЯ-текстов и морфологических анализаторов, посвящен раздел 1.3. Объектно-предикатные системы как составляющая концептуально-функциональной модели и средства формального описания семантики, в частности, значений морфем, рассматриваются в разделе 1.4.

В результате анализа работ в области разработки систем обработки знаний получены следующие Выводы по главе 1.

1. При рассмотрении тенденций развития систем и технологий обработки ЕЯ-текстов исследователи, как правило, придерживаются сходной стратегической линии и выделяют три подхода, отмеченные М.Г.Мальковским как лингвистический, экспериментальный и прагматический. Лингвистический подход характеризуется стремлением к использованию в ЕЯ-системах максимально полных моделей языка, к построению полностью явных, эксплицитных, описаний и к определению максимально адекватной общей структуры этих описаний. Экспериментальный подход к построению лингвистических систем, как правило, используется специалистами в области искусственного интеллекта, и вызван глобальностью и сложностью решаемых

задач, необходимостью сокращения проблемной области и языка текста, попыткой глубокого проникновения в содержание ЕЯ-текста и рассматриваемого контекста, стремлением построить реально функционирующую эффективную систему. *Прагматические системы*, как правило, ассоциируются с реальными системами, созданными для решения конкретного круга задач в узкой проблемной области, и достигающих высокой эффективности за счет ее ограниченности, однозначности контекста и примитивности языкового интерфейса.

2. Идеи А.С.Нариньяни о модельном, а не алгоритмическом подходе, о децентрализованном и асинхронном анализе текстов, об организованном сообществе активных *constraint-based агентов* - а это в нашем определении *концептуально-функциональные лингвистические модели*, весьма четко и убедительно очерчивают перспективные направления развития систем и технологий обработки информации.

3. Принципы построения лингвистической модели и реализации функций лингвистического процессора естественным образом находятся в прямой зависимости от прагматической ориентированности лингвистической модели и специфики соответствующего класса ЛП. Выявление и учет этих прагматических характеристик дает возможность строить более эффективные системы, ввиду применения адекватных методик, ориентированных на узкий круг задач.

4. В ряде проанализированных работ авторы явно или неявно высказывают или даже используют при разработке конкретных систем идеи, близкие по содержанию к идее прагматически-ориентированного подхода. Однако прагматика даже в узком ее понимании по-прежнему представляет собой наименее разработанный аспект систем обработки знаний, а для более общих случаев построения прагматически-ориентированных моделей и технологий вопрос в такой постановке практически в литературе не рассматривается.

5. Нерешенность задач удовлетворительного дружественного естественно-языкового диалога, машинного перевода с языка на язык являются косвенным подтверждением правильности прагматически-ориентированного подхода, когда исследуются и находятся такие проблемные области, для которых удастся построить вполне удовлетворительные лингвистические модели, т.е. объединить в рамках единой технологии концептуально-формальные средства, достаточные с методологической и практической точек зрения.

6. Очевидно, выделение классов предикатов и объектов в "модели мира" есть процесс перманентный, требующий глубокой лингвистической интуиции от исследователя. Ни одна из рассмотренных классификаций предикатов и объектов не является полной и завершенной и навряд ли вызовет сомнение у исследователей утверждение, что вопросы полноты и достаточности объектно-предикатной системы могут решаться лишь в ходе практического ее использования. Следовательно, весьма актуально иметь некий инструментарий для фиксирования выделенных объектов и отношений, а также

автоматизированного поиска и установления их в огромных массивах машиночитаемых ЕЯ-текстов. Таким инструментом является структурно-функциональная модель татарских морфем, описываемая в главе 2 диссертации.

7. Диалоговое взаимодействие пользователя с автоматизированной системой протекает в одном из следующих режимов: 1) активна система, когда на вопросы системы отвечает пользователь, 2) активен пользователь, когда на запрос пользователя определенным образом реагирует система, 3) двусторонне активный диалог, когда пользователь и система меняются ролями в ходе общения.

8. Примером диалоговой модели, наиболее естественно моделирующей вопросно-ответную ситуацию, т.е. режим, когда активна система и пассивен пользователь, является вопросно-ответный диалог в автоматизированной обучающей системе. Вопросно-ответная ситуация в автоматизированных обучающих системах имеет свои особенности (специфика входного текста и формальной основы анализа, глубина лексико-семантического анализа и др.), учет которых позволяет строить прагматически-ориентированные лингвистические модели как основы эффективных анализаторов ответов обучаемого.

9. Традиционные средства анализа по ключевым параметрам и эталонам ответов и универсальные методы контроля знаний в АОС не позволяют анализировать естественно-языковые тексты и не способствуют смысловому анализу ответа обучаемого. Требуется расширить возможности контроля ответов обучаемого лингвистическими процедурами, позволяющими анализировать ЕЯ-тексты по смыслу.

10. Сила унифицированных грамматических формализмов, как основы семантического анализа ЕЯ-текстов, в преимуществе их возможностей для разработки прагматически-ориентированных грамматик. Как показывает опыт исследователей, большие грамматики могут быть описаны, но их реализация чрезвычайно сложна, практически, нереальна. В настоящее время не существует удовлетворительных методов для эффективной реализации подобных грамматик.

11. Весьма продуктивным при разработке прагматически-ориентированных когнитивных и диалоговых моделей признается подход к кодированию семантической информации через типовые конструкции. Это обеспечивает автоматическую конвертацию выражений, записанных на естественном языке, в выражения на языке описания семантики. Такой подход является близкой вариацией принципа разработки индивидуальных концептуальных грамматик на основе семантической классификации вопросно-ответных текстов, раскрываемого в диссертации.

12. Важной задачей, являющейся частью проблемы формализации высказывания, но в настоящее время не имеющей удовлетворительного решения, является задача сегментации, т.е. разбиение текста на части, порции, элементы, являющиеся конструкциями, из которых складываются более сложные структуры.

13. Обзоры, касающиеся концептуально-формальных грамматик, сконцентрированы на разработках, которые относятся к морфологическим анализаторам, широко используемым сегодня на практике, при описании которых остаются в стороне многие теоретические результаты, особенности конкретных языков, учет которых, безусловно, позволяет использовать более адекватные технологии и повышает эффективность лингвопроцессоров. Следующие две проблемы практически являются базовыми в исследуемой области: 1) выявление и учет морфологических альтернатив: - многозначности морфем, зависимости их от контекста; 2) формальное описание морфотактики: - правил комбинирования основ и аффиксов по определенным правилам, которые должны быть описаны в модели. Это особенно важно для агглютинативных языков с богатой морфологией, к которым относится и татарский язык, рассматриваемый в качестве объекта исследования и демонстрации в диссертации.

14. Большинство моделей корректоров реализует следующий сходный алгоритм. На основе исходного словаря основ порождается список словоформ, далее этот список некоторым образом упаковывается и в нем реализуется функция поиска. Такой алгоритм называется парадигматическим. Распространенным для решения проблемы морфологических альтернатив (определение корректности соответствующего аффикса) является применение метода *cut-and-paste*, заключающегося определения правильной формы путем удаления и присоединения букв к концу слова. Именно такой подход, являющийся частью генеративного алгоритма, использован нами при построении татарского морфологического корректора. Третий подход, используемый для разработки концептуально-формальных морфологических моделей - использование технологии конечного состояния для автоматического распознавания и генерации словоформ. Он основан на замечании, что правила морфологических альтернатив могут быть реализованы трансдюсерами конечного состояния, а также на идее, что возможные комбинации основ и аффиксов могут быть кодированы как сеть конечного состояния. На таком подходе - использовании двухуровневой морфологической модели, устроен татарский морфологический анализатор, реализованный в среде программного инструментария PC-KIMMO.

В главе 2 описывается открытая структурно-функциональная модель морфем и интегрированная программная среда для заполнения, модификации и эксплуатации заполненной модели и технология применения модели для обработки естественно-языковой информации, а также формальный инструментарий для описания значений аффиксальных морфем. В разделе 2.1 приводится краткий анализ проблемы. Как известно, татарский язык как язык агглютинативного типа обладает морфологией, богатой по составу, сложной по содержанию, но простой, регулярной и закономерной по структуре. В связи с этим актуальной является задача построения прагматически-ориентированной компьютерной модели морфем, позволяющей провести полную "инвентаризацию" татарских аффиксальных морфем с описанием функций и ситуаций проявления их в словоформе.

База неформального описания аффиксальных морфем, осуществленного Ф.А.Ганиевым, Ф.С. Сафиуллиной и Л.Х.Ризвановой, является важным и полезным упорядоченным первичным лингвистическим материалом, упрощающим этап начального заполнения базы модели и позволяющим обратить усилия на концептуальные (полнота описания характеристик), технологические (удобство обработки, прагматичность), инструментальные аспекты разработки и реализации (открытость, наличие функций и средств для доступа и применения).

Из названия самой модели - структурно-функциональная, следует, во-первых, что она структурно, т.е. на разных языковых уровнях (фонологическом, морфологическом, синтаксическом, семантическом), отображает проявление свойств морфемы. Модель является открытой, что позволяет вносить в нее даже единичные, незначимые на первый взгляд, характеристики или проявления аффикса в словоформе, при необходимости модифицируя и саму структуру. Во-вторых, модель функциональная, т.е. после наполнения ее лингвистами-экспертами является информационно-справочной и учебно-консультационной базой при изучении и преподавании татарского языка, а также мета-моделью, используемой для разработки прагматически-ориентированных лингвистических моделей в качестве формальной базы для лингвопроцессоров. Еще одно перспективное применение подобных моделей открывается в рамках "модельной" технологии анализа ЕЯ-текстов, предложенной А.С.Нариньяни, и описанной нами в главе 1.

В разделе 2.2 раскрывается структурно-функциональная модель аффиксальных морфем как концептуально-функциональное описание элементов морфологии.

Модель имеет структуру, в которой каждая морфема характеризуется по восьми основным параметрам, называемым аспектами. Каждый основной параметр подразделяется на подпараметры, а те, в свою очередь, также могут подразделяться на подпараметры и т.д. В разделе 2.2.1 приводится описание каждого аспекта модели.

1. Функциональный аспект — это признаки, характеристики морфем как грамматической единицы.

Морфо-функциональный аспект - это признак того, является ли морфема свободной или же связанной (аналитической или синтетической). Например, -Дан (аблатив, исходный падеж) — синтетическая морфема:

урам+Дан — урамнан 'улица'+[аблатив] — 'с улицы'

Синтактико-функциональный аспект указывает, к какой из трех категорий — словоизменительной, словообразовательной или модальной принадлежит описываемая морфема. Например, морфема -Дан (аблатив, исходный падеж) является словоизменительной, поскольку словоформа, получаемая из исходной с помощью присоединения данной морфемы, сохраняет свое лексическое значение.

Семантико-функциональный аспект показывает, обладает ли данная морфема функциональным значением (т.е. для чего? для выражения чего?).

2. Морфологический аспект.

Порядок следования морфем в словоформе. Данный аспект позволяет наблюдать, какие морфемы могут находиться слева и справа (иначе, являются ближайшими соседями, окружением). Пример реализации, когда исследуются морфемы, следующие непосредственно справа от *-Дан*:

№	Морфем а	Следует справа (+/-)	Комментарий	Пример
1.	[]	+	нулевой аффикс	<i>авыл-дан</i> (<i>'из деревни'</i>)
2.	<i>-ГА</i>	—	Направительн. Падеж	
3.	<i>-ДЫр</i>	+	модальность сомнение/ уверённость	<i>авыл-дан-дыр</i> (<i>'наверное, из деревни'</i>) (<i>'точно, из деревни'</i>)

В тех случаях, когда для различных категорий одной части речи наблюдаются различные морфологические свойства, часть речи также, в свою очередь, разбивается на подкатегории и, соответственно, появляются новые подкатегории в модели. В разделе *«Морфологический аспект. Морфемы, находящиеся справа от исследуемой»* рассматриваются все послелогии и послеложные слова, приведенные в академической грамматике татарского языка.

Рекурсия. Параметр рекурсия показывает, наблюдается ли для рассматриваемой морфемы явление рекурсии, т.е. может ли данная морфема встречаться в словоформе больше одного раза, формируя новый смысл по той же схеме. Примеры с аффиксом *-ДАгы*:

<i>авыл+дагы</i> (<i>'те, кто в деревне'</i>)	<i>'деревня'+падеж: место-временной-2,</i> (<i>предложный, указание на нечто</i>),
<i>авыл+дагы+лар</i> (<i>'те, кто в деревне'</i>)	<i>'деревня'+падеж: место-временной-2,</i> (<i>предложный, указание на нечто</i>)+мн.ч.,
<i>авыл-дагы-лар-дагы</i> (<i>'то, что у тех, кто в деревне'</i>);	<i>'деревня'+падеж: место-временной-2,</i> (<i>предложный, указание на нечто</i>)+мн.ч. + падеж: <i>место-временной-2, предложный (указание на нечто).</i>

3. Синтаксический аспект.

Порядок следования морфем в тексте. Данный аспект показывает, что именно (т.е. какая корневая морфема) может находиться слева и справа от словоформы, когда рассматриваемая аффиксальная морфема является самой правой в словоформе.

Фразеологические словосочетания. Примеры фразеологических словосочетаний для морфемы *-ГА*: *акыл+га килу* (*'одуматься'*, буквально: *'прийти в ум'*), *йокы+га киту* (*'заснуть'*, буквально: *'уйти в сон'*).

Редукция. Данный аспект показывает, наблюдается ли явление пропуска рассматриваемой морфемы в предложениях с однородными членами предложения, с сохранением ее в последнем члене. В лингвистике это явление называется аффиксальным эллипсисом. Например, следующие фразы являются идентичными: 1) *Без урманнардан, болыннардан, таулардан үттек. 'Мы лес'+мн.ч.+исх.падеж, 'луг'+мн.ч.+исх.п., 'гора'+мн.ч.+исх.п. 'прошли' ('Мы прошли сквозь леса, сквозь луга, сквозь горы').* 2) *Без урманнар, болыннар, таулардан үттек. ('Мы прошли сквозь леса, луга, горы').* Таким образом, аффикс *-Дан* обладает свойством редукции. В данном примере свойством редукции обладает и служебное слово *сквозь* (в русском варианте), т.е. оно выводится за последовательность перечисляемых слов без изменения их значений. В татарском языке свойство редукции допускает несколько итераций, т.е. за последовательность (так сказать, “за скобки”) могут быть вынесены и другие предыдущие аффиксы, являющиеся общими для перечисляемых словоформ. В рассмотренном примере это будет выглядеть следующим образом: *Без урман, болын, таулардан үттек 'Мы лес, луг, гора'+мн.ч.+исх.п. прошли'.*

4. Семантический аспект.

Тождественная лексическая семантика. Указывает, изменяет лексическую семантику рассматриваемая морфема или нет.

Значение. Указывает на значение, приобретаемое словоформой после присоединения данной морфемы. Данный параметр можно назвать контекстным, поскольку значение слова зависит как от параметров аффикса, так и от окружения, в котором оно встречается. На основе обобщения описаний тематической структуры, приведенной в работе М.З.Закиева, заполнены следующие подпараметры Лексико-семантического аспекта, описывающие значения морфемы на примере аффикса *-ГА*.

4.2.1. Исполнитель процесса	+	укытучыга хат укыту (<i>'дать учителю прочитать письмо'</i>), т.е. учитель читает письмо
4.2.2. Объект		
4.2.2.1. Прямой	—	
4.2.2.2. Косвенный	+	укучыга әйтү (<i>'сказать ученику'</i>)
4.2.3. Цель процесса	+	утынга бару (<i>'ехать за дровами'</i>)
4.2.4. Причина совершения процесса	+	матурлыкка соклану (<i>'восхищаться красотой'</i>)
4.2.5. Время совершения процесса	+	киләсе атнага кайту (<i>'вернуться на следующей неделе'</i>)
4.2.6. Место совершения процесса		
4.2.6.1. Конечный пункт	+	урманга бару

назначения		(идти в лес')
4.2.6.2. Исходный пункт	—	
4.2.7. Способ совершения процесса	+	бәтен көчкә чабу (бежать изо-всех сил')
4.2.8. Мера, степень процесса, действия		
4.2.8.1. Мера процесса	+	бер төнгә кайту (вернуться на ночь')
4.2.8.2. Мера объекта процесса	+	ун сумга алу (купить на десять рублей').

Подробному рассмотрению семантического аспекта посвящен раздел 2.3.

5. Синонимия. Параметр содержит морфемы или словоформы, синонимичные рассматриваемой морфеме. Морфемы, тождественные по смыслу, но различающиеся форме, называются синонимичными.

6. Омонимия. Ситуация противоположная синонимии, когда наблюдается полная тождественность состава морфем и их порядка при различных значениях.

7. Антонимия. Содержит морфемы, которые при присоединении к словоформе придают ей противоположный смысл. Например, для аффикса -лы/ (обладательный падеж) антонимом является аффикс -сыз (лишительный падеж):

тоз+лы ('солёный') - тоз+сыз ('не солёный')
 соль+обладат.падеж соль+лишит.падеж.

8. Таблица алломорфов. Морфы, обладающие тождеством значения и фонематической близостью, в которых фонематические чередования обусловлены позиционно, т.е. соседними морфемами, называются алломорфами морфем. В данной таблице представляются все алломорфы описываемой морфемы, являющиеся поверхностными представлениями соответствующей морфемы в некотором контексте по правилам фонологии.

В силу открытости функционально-структурной модели татарских морфем, при заполнении конкретных экземпляров морфем она может дополняться как новыми аспектами, так и новыми подпараметрами, характеризующими данную морфему. Очевидно, что такие характеристики, как адекватность модели реальным языковым явлениям, полнота описания конкретных морфем, могут оцениваться только исходя из устойчивости модели, ее корректного функционирования при решении практических задач на основе этой модели.

Попытка описать значения каждой морфемы в подразделе "Семантический аспект" в некотором унифицированном виде, с целью их дальнейшего использования в лингвопроцессорах, привела к необходимости систематизации отношений, выражаемых аффиксальными морфемами, и создания целостной картины этих отношений. Известно, что значение морфемы зачастую может быть установлено только исходя из контекста, в котором она встречается и является трудно формализуемым параметром. В

этой главе приводится описание объектно-предикатной системы, используемой в семантическом аспекте структурно-функциональной модели для формального представления значений морфем. Формальное описание значений аффиксальных морфем преследует цели, определяемые следующими прагматическими установками: 1) интерпретация некоторого фрагмента реального мира (заданного априори); 2) применение модели для решения определенных задач, например, формального представления значений аффиксальных морфем, для использования их в учебных задачах или как базис для задач машинного перевода.

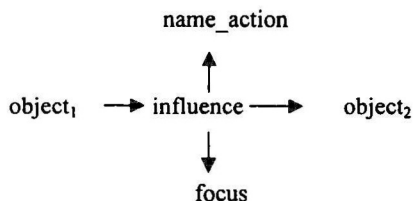
Аффиксальные морфемы по определению как минимальные значащие единицы языка имеют хотя бы одно значение, проявляющееся при использовании его в словоформе. В татарском языке, как правило, в зависимости от окружения, аффиксальные морфемы имеют различные интерпретации, точнее, используются для поверхностного оформления различного рода контекстов, следовательно, обладают различными значениями.

В разделе 2.3.2 подробно рассматривается отображение некоторых контекстов, в формировании которых участвует морфема *-ГА*. Здесь приведем два примера описания значения *-ГА* средствами объектно-предикатной системы. Полное описание структурно-функциональной модели для аффиксальной морфемы *-ГА* приводится в Приложении 4.

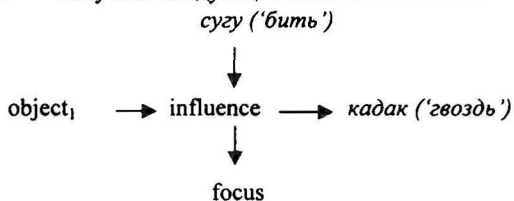
Значение 1. *Концептуально: зависимое слово с аффиксом -ГА выражает косвенный объект процесса.*

В этом контексте словоформы с морфемой *-ГА* представляют собой, либо объект физического или информационного воздействия, либо объект, который становится владельцем некоторого объекта.

Физическое воздействие представляется следующим сценарием:



При подстановке конкретного словосочетания: *кадакка сугу 'ударить по звезде'* — получаем следующее заполнение схемы:

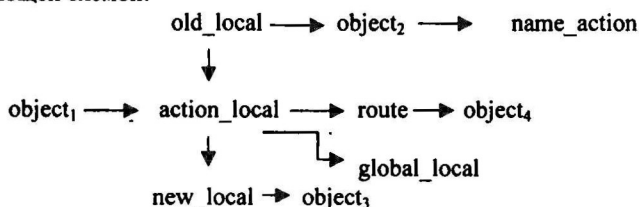


В объектно-предикатной системе рассматриваемая словоформа заполняет следующую ролевую ситуацию: *Отношение Действие: Действие-воздействие: Объект воздействия.*

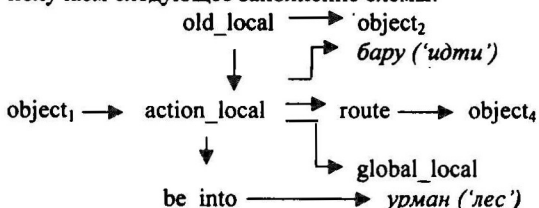
Значение 2. Концептуально: зависимое слово с аффиксом -ГА выражает конечный пункт направления процесса, обозначенного глаголом движения.

В данном контексте выражается конечный пункт действия по изменению пространственного отношения.

Действие по изменению пространственного отношения представляется следующей схемой:



При подстановке конкретного словосочетания: *урманга бару 'идти в лес'* — получаем следующее заполнение схемы:



В объектно-предикатной системе рассматриваемая словоформа заполняет следующую ролевую ситуацию: *Отношение Действие: Действие по изменению пространственного отношения: Новое местоположение.*

Интегрированная среда для создания, модификации и применения модели морфем описывается в разделе 2.4. На основе структурно-функциональной прагматически-ориентированной модели морфем, описанной в п.п.2.2-2.3, разработана база данных и создан программный инструментарий, позволяющий обрабатывать ее. В настоящее время разработан и реализован программный инструментарий в среде Windows'95 для формирования, заполнения и модификации модели в интерактивном режиме, заполнена база аффиксальных морфем и осуществляется заполнение модели соответствующими конкретными характеристиками морфем. На основе модели морфем разработаны программа лемматизации, т.е. выделения основы в произвольной словоформе, а также программа разбиения слова на слоги, весьма актуальная для программы синтеза татарской речи на слогово-фонологической основе и для издательских технологий.

В Выводах отмечается, что особенно важным и в корне отличающим структурно-функциональную модель от электронных словарей, тем более от некомпьютеризованных описаний морфем, является возможность ввода

логических запросов. Например, заполненная модель морфем с блоком логического вывода практически “извечный” спор о количестве падежей в татарском языке превращает в чисто вычислительный процесс.

Формальный аппарат, описанный в данной главе, позволяет максимально полно отразить те отношения, в лексикализации которых участвуют аффиксальные морфемы, повышает репрезентативность, прозрачность и прикладную ценность модели.

Глава 3 посвящена описанию базовых принципов построения и архитектуры системы обработки информации в вопросно-ответном контексте как пример ЛП естественно-языковых текстов на основе прагматически-ориентированной модели, относящейся к классу диалоговых моделей.

В разделе 3.1 описываются прагматические особенности, теоретические положения и понятийный аппарат для описания соответствующей модели, принципы построения.

Построение системы обработки текстов на ЕЯ в контексте, управляемом вопросом системы к пользователю, имеет специфику, выгодно отличающую ее от других ЕЯ-диалоговых систем и создающую реальные предпосылки для построения эффективного смыслового интерпретатора. Это предопределено, прежде всего, следующими особенностями вопросно-ответной ситуации.

Особенность входного текста и его структуры.

Входной текст - это ответ на заданный вопрос с ожидаемым смыслом. Контекст диалога управляется вопросом системы и накладывает определенные ограничения как на форму ответа, так и на его содержание.

Особенность лингвистической базы для анализа ответа.

В силу активности системы, задающей вопрос, ей заранее известен контекст, в котором будет происходить интерпретация ответа и достаточно легко может быть сформирована модель текста, адекватная ожидаемому как по лексике, так и по форме изложения.

Система обработки ЕЯ-текстов в контексте, управляемом вопросом, включает базу знаний (*экземпляры фреймов ответов и индивидуальные концептуальные грамматики*), лексический процессор и семантический интерпретатор. Интерпретация ответного текста происходит следующим образом. Текст поступает в лексический процессор и на основе экземпляров фреймов переводится в канонизированное представление в виде последовательности специальных смыслообразующих единиц. Часть нераспознанной информации на лексическом уровне может представлять интерес для дальнейшего разбора (например, для проверки на непротиворечивость с ожидаемой частью ответа), поэтому накапливается в специальных файлах. Далее канонический текст интерпретируется с привлечением специальных семантических схем - *индивидуальных концептуальных грамматик*. Результат формируется в виде некоторого вектора ситуации (иначе, диагностики ответа), по которому принимается управляющее действие системой.

Построение лингвистического процессора базируется на следующих двух методологических принципах и шести принципах реализации прагматически-ориентированной модели.

Методологические принципы:

Принцип детерминированности контекста. В силу активности система "погружает" пользователя в некий контекст, который достаточно строго определяется заданным вопросом. Соответственно, содержание ответа, его лексикон и даже форма и, отчасти, объем предопределены и пользователь с необходимостью отвечает на вопрос в управляемых рамках.

Принцип ожидаемости смысла ответа. По заданному вопросу система знает пространство значений вопроса, т.е. ей заранее известен контекст, в котором будет происходить интерпретация ответа и достаточно легко может быть сформирована модель текста, адекватная ожидаемому ответу как по лексике, так по форме изложения и семантической конструкции.

Принцип ожидаемого смысла является справедливым и для любого разумного диалога, в котором обмен информацией создает общий контекст, определяющий дальнейшее его протекание. В этой связи исследование принципов ожидаемого смысла и настройки системы на заданный контекст представляется естественным и для выявления механизмов двусторонне-активного диалога.

Принципы реализации:

Принцип 1. Выделение системы смыслообразующих единиц (концептул) с целью трансформации проблемы семантического анализа вопросно-ответного текста в проблему синтаксического анализа в условиях использования детерминирующей роли контекста.

Принцип выделения концептул приводит к необходимости провести типизацию понятий, отношений, грамматических признаков и специальных ролей лексем и установления соответствия между ними и концептулами в управляемом контексте, т.е. в контексте заданного вопроса.

Каждое осмысленное предложение ПО можно перевести в текст, составленный из типов понятий и типов отношений, без детального учета грамматических особенностей ЕЯ, отображая каждое понятие или отношение в соответствующую концептулу. Этот процесс будем называть канонизацией текста, а возникающий при этом обобщенный смысл, характеризующий взаимодействие концептул, - каноническим смыслом исходного ЕЯ-текста. Принцип выделения концептул позволяет провести типизацию понятий, отношений, грамматических признаков и специальных ролей лексем. Выделение концептул производится на основе анализа типов лексем и их ролей в вопросно-ответных текстах.

Принцип 2. Семантическая классификация вопросно-ответных текстов на основе типовых отношений: - выделение конкретных типов отношений, типов вопросов и классов ответов для реализации этой детерминирующей роли контекста.

В вопросно-ответном диалоге система функционирует в условиях определенного контекста и она способна четко очертить круг ожидаемых

возможных ответов, т.е. значений вопроса, и декодировать ожидаемый смысл из многообразия грамматически правильно построенных фраз в соответствии с этим предварительным знанием. Смысловая типизация вопросов и семантическая классификация значений вопроса дают возможность противопоставить каждому типу вопроса ограниченный набор допустимых ответных формул, т.е. логически правильных смысловых конструкций. Можно рассматривать совокупность этих формул, соответствующих конкретному типу вопроса, как некоторую грамматику, кодирующую конструкции, передающие правильный смысл ответа.

Принадлежность ответа к некоторому классу ответов определяется не по его объему и содержанию, и не по форме вопроса, а по типу вопроса системы и по ожидаемому смыслу.

Принцип 3. Разработка ИКГ семантических классов, отражающих смысловые конструкции ответов соответствующих классов и в совокупности составляющих концептуальную грамматику как схему реализации принципа трансформации семантики в синтаксис, служащей формальной основой для построения семантического интерпретатора, ориентированного на "слушающего".

Сочетания понятий и отношений в текстах, соответствующих определенным семантическим классам, имеют достаточно устойчивые частотные характеристики. Следовательно, при создании системы семантической интерпретации логично ожидать в анализируемом тексте семантические конструкции, имеющие наиболее высокие частотные характеристики для рассматриваемого контекста.

Принцип 4. Сегментация вопросно-ответных текстов по минимальным смысловым конструкциям для рекурсивного применения правил концептуальной грамматики (базовых смысловых формул).

В проблематике семантического анализа текстов на ЕЯ, особенно для практической реализации разработок, оказывается важной задача членения входного текста на такие части, к которым рекурсивно применимы простые формулы. Сложный текст, представляет собой линейную и/или иерархическую последовательность смысловых частей, относящихся к тому или иному семантическому классу ответов. *Часть сложного текста, или полный текст, относящийся к определенному семантическому классу, названа сегментом.* Таким образом, сложный текст является линейно и /или иерархически организованным множеством сегментов.

В известных системах понимания ЕЯ практически отсутствуют эффективные механизмы выделения сегментов в анализируемом связном тексте для применения к ним ограниченного набора унифицированных правил анализа. Глубинные причины такого положения лежат в сложности самой проблемы членения входного текста на соответствующие смысловые части. Это посылить только действительно интеллектуальной системе, способной на основе плавающего (уточняющего смысл части текста по месту чтения) контекста выделять смысловые конструкции, рекурсивно идентифицируемые с правилами ИКГ соответствующих классов ответов.

Для реализации принципа сегментации важно ответить на следующие два вопроса: как определить контекст, в рамках которого входной текст должен анализироваться на смысловую корректность, и каким образом выделять в тексте сегменты, чтобы к ним были рекурсивно применимы грамматические формулы.

Ответ на первый вопрос ведет нас, в общем случае, к необходимости подробного анализа проблем плавающего контекста, условий его изменения, сохраняющих смысловую непрерывность восприятия и ряда других методологических, лингвистических, психологических и семиотических проблем. В нашем случае мы сознательно идем на некоторое упрощение ситуации, фиксируя контекст по заданному вопросу. В силу этого входной текст, т.е. ответ обучаемого, однозначно попадает в рассматриваемый контекст и фактически содержит ожидаемый смысл.

Для применения соответствующих ИКГ требуется определить, к какому семантическому классу ответов относится вводимый текст. В случае вопросно-ответного текста система способна заранее по задаваемому вопросу предопределить семантический класс ожидаемого ответа, тем самым, предопределяя и соответствующую ИКГ, применяемую для его смыслового анализа. Семантическая классификация вопросно-ответных текстов производится от простого к сложному. Вначале определяются простые семантические классы ответов, т.е. ответы, в которых раскрывается смысл вида "понятие-отношение-понятие". Затем из таких конструкций строятся более сложные семантические классы, представляющие собой комбинации простых классов, как линейные, так и иерархические, отражающие существование связанных текстов из простых, сложносочиненных и сложноподчиненных предложений. Соответственно, сегментация текстов приводит к построению как линейных, так и иерархических представлений, которые рекурсивно распознаются на основе определенных ИКГ.

Принцип 5. *Релевантность представления знаний (модели ответа) по смысловой структуре и лексическому наполнению ожидаемому ответному тексту.*

Очевидно, наиболее эффективный диалог, т.е. достаточно адекватная и реактивная интерпретация. входного текста будет осуществляться при соблюдении принципа релевантности представления знаний (модели ответа) по смысловой структуре и лексическому наполнению ожидаемому ответному тексту. Это является естественным требованием к системе интерпретации, моделирующей человеко-машинный диалог, т.к. в управляемом контексте активный участник диалога всегда имеет возможность заранее построить модель ответа адекватно ожидаемому контексту по своему вопросу.

Принцип 6. *Принцип открытости системы, обеспечивающий развитие системы путем накопления новых знаний на основе устойчивых статистических характеристик, в том числе, путем расширения множества обобщенных семантических единиц, введения новых типов вопросов и классов ответов, сортировки и расширения правил КГ, введения новых ИКГ.*

В разделе 3.2.1 описываются концептулы, отражающие типы понятий.

Обозначим через K_S множество концептуал, отражающих различные типы понятий, т.е.

$K_S = \{SS, SS(i), SO, S_{оп}, SA, SP\}$, где

SS - концептула, отражающая главное понятие (первая буква S - признак того, что концептула отражает понятие). Главное понятие - это понятие, относительно которого задан вопрос (т.е. понятие, раскрываемое в данном контексте).

Сложные тексты могут содержать несколько понятий, связи которых раскрываются в анализируемых предложениях, каждый из которых в процессе анализа определенной части предложения может, в свою очередь, выступать в роли главного понятия. Для их различения в пределах анализируемого текста вводится обозначение:

$SS(i)$ - концептула, отражающая i -е главное понятие.

SO - концептула, отражающая понятие, состоящее в некотором определенном отношении с главным понятием.

SA - концептула, отражающая понятие-аргумент.

SP - концептула, отражающая понятие-результат.

$S_{оп}$ - концептула, отражающая обобщенное понятие, т.е. понятие, находящееся по отношению к главному на более высоком уровне в иерархии понятий предметной области.

В разделе 3.2.2 вводятся понятия *Типового Отношения (ТО)* как обобщенного отношения, объединяющего элементарные отношения в семейства отношений, отражающих однотипный смысл и раскрывающих определенный признак понятий предметной области. Например, элементарные отношения, выраженные лексемами типа 'играет', 'спит', 'плавает', объединяются в семейство *ТО СОСТОЯНИЕ* по признаку: *выражать состояние понятия (т.е. отражать однотипный смысл, раскрывающий признак понятия ПО - находиться в некотором состоянии)*.

Аналогично определяются *ТО ДЕЙСТВИЕ*, *СОСТАВ ВКЛЮЧЕНИЕ*, *ВРО* (Временное Отношение), *ПРО* (Пространственное Отношение), *ТО КЛО* (Количественное Отношение), *КЧО* (Качественное Отношение). Приведенный список *ТО* подробно описан в диссертации и может быть расширен в зависимости от *ПО* по усмотрению АВТОРа. Также вводится понятие *Составного Отношения (СО)* как *устойчивая комбинация двух ТО ДЕЙСТВИЕ: действий, направленных на аргумент (т.е. отношение SS к SA) и результат (т.е. отношение SS к SP), соответственно*. В следующем тексте раскрывается *СО ФУНКЦИЯ* понятия S_1 : ' S_1 переводит S_2 в S_3 '.

В *ПО* возможны тексты с другими достаточно устойчивыми связками *ТО*. В целях повышения эффективности анализа целесообразно рассматривать такие тексты как ответы, раскрывающие *СО*. Например, рассмотрим следующие тексты:

- 1) *Файл курса хранится на диске.*
- 2) *Деталь вращается вокруг оси.*
- 3) *Компиляция выполняется раньше загрузки.*
- 4) *Сказуемое в английском предложении стоит после подлежащего.*

В предложении (1) раскрываются два отношения: 'Файл курса хранится' (ТО СОСТОЯНИЕ) и 'Файл курса на диске' (ТО ПРО).

В предложении (2) раскрываются также и ТО ПРО: 'Деталь вращается' (ТО СОСТОЯНИЕ) и 'Деталь вокруг оси' (ТО ПРО).

В предложении (3) раскрываются ТО СОСТОЯНИЕ и ТО ВРО: 'Компиляция выполняется' и 'раньше загрузки', соответственно.

В предложении (4) раскрываются ТО ПРО, СОСТОЯНИЕ, ВРО: "Сказуемое в английском предложении" (ТО ПРО), "Сказуемое стоит" (ТО СОСТОЯНИЕ), "Сказуемое после подлежащего" (ТО ВРО).

В зависимости от ПО и по усмотрению АВТОРа тексты типа (1) и (2) могут рассматриваться как ответы, раскрывающие одно СО СОСТОЯНИЕ-ПРО; текст типа (3) - как ответ, раскрывающий СО СОСТОЯНИЕ-ВРО; текст типа (4) - как ответ, раскрывающий СО ПРО-СОСТОЯНИЕ-ВРО.

Введем формальные обозначения для концептул соответствующих типов отношений (первая буква R означает, что концептула отражает отношение). R_C - это концептула, соответствующая ТО СОСТОЯНИЕ; $R_{СОСТ}$ - ТО СОСТАВ; $R_{ВКЛ}$ - ТО ВКЛЮЧЕНИЕ; R_D - ТО ДЕЙСТВИЕ; $R_{ВРО}$ - ТО ВРО; $R_{ПРО}$ - ТО ПРО; $R_{КЛО}$ - ТО КЛО; $R_{КЧО}$ - ТО КЧО; R_{SO} - концептула, отражающая отношение SS к SO; R_{OS} - концептула, отражающая отношение SO к SS; R_A - концептула, отражающая отношение SS к SA; R_P - концептула, отражающая отношение SS к SP.

Через K_R обозначено множество концептул, отражающих различные типы отношений, т.е. $K_R = \{R_C, R_{СОСТ}, R_{ВКЛ}, R_D, R_{ВРО}, R_{ПРО}, R_{КЛО}, R_{КЧО}, R_{SO}, R_{OS}, R_A, R_P\}$. Здесь R_{SO} , R_{OS} принимают значения из следующего множества: $\{R_C, R_{СОСТ}, R_{ВКЛ}, R_D, R_{ВРО}, R_{ПРО}, R_{КЛО}, R_{КЧО}\}$.

Пояснения. Пары концептул (R_{SO} , R_{OS}) и (R_A , R_P) являются взаимоисключающими в канонизированном тексте ответа. Взаимоисключающими являются также концептулы R_{SO} и R_{OS} . Хотя бы одно из отношений, выраженных концептулами R_A и R_P , содержится в ответе явно, а другое имеет определенное место и присутствует в ответе неявно, т.е. в контексте предложения. Например, в тексте '*S1 переводит S2 в S3*' - понятие *S1* есть SS, понятие *S2* - SA и понятие *S3* - SP.

В этом предложении раскрываются два отношения типа ДЕЙСТВИЕ: '*S1 переводит S2*' и '*S1 получает S3*', связанные единым смыслом - 'раскрыть функцию понятия *S1*'. Здесь R_A - 'переводит' - в тексте задано явно, R_P - 'получает' - задано неявно. Таким образом, концептулы первой группы (обозначим K_1) включают следующее множество прагматических ролей (концептул): $K_1 = K_S \cup K_R$.

Полный отказ от элементов классической грамматики ЕЯ оправдан не во всех случаях. Иногда смысл предложения определяется предлогами, падежными окончаниями слов и др., и их учет позволяет существенно сократить пространство возможных схем разбора текста, соответственно, упрощая семантический анализ. Поэтому вводится дополнительная семантическая единица - грамматическая роль лексем или их частей для указания соответствующих грамматических признаков естественного языка,

например, *GPr* - концептула, отражающая предлог перед *SP* (первая буква *G* означает, что концептула отражает грамматическую роль), например, предлоги 'в', 'на', 'к'.

Смысл анализируемого текста зависит также от специфики проблемной области. Этим вызвано введение третьего типа концептул - специальных ролей лексем в ответе пользователя, например, *LN* - концептула, отражающая необязательную лексему, т.е. лексема, отсутствие или наличие которой в ответе не влияет на смысл ответа (первая буква *L* означает, что концептула отражает специальную роль лексемы).

В общем случае канонический смысл текста определяется сочетанием концептул всех указанных типов. На основании введенной классификации *ТО* и *СО* производится семантическая классификация текстов: *вопросов АВТОРА и ответов обучаемого*.

На форму задания вопросов не накладываются специальных ограничений. Ограничения естественным образом исходят лишь из требования, что вопрос должен быть однозначно понят обучаемым (т.е. по тексту вопроса должно быть понятно, раскрытие какого смысла требуется в ответе). Так, выделяются следующие *типы вопросов и соответствующие классы ответов*.

I. Вопросы, требующие явного задания в ответе ключевых понятий (без учета отношений между понятиями; отношения явно заданы в вопросе). Сюда относятся вопросы: *'Напишите программу вычисления функции на Паскале', 'Назовите состав компилятора'*.

Этому типу вопросов соответствуют классы ответов, в которых обязательно явно содержатся *ключевые понятия*. Например, *ответы выборочного типа* (даны несколько ответов, необходимо указать правильный ответ); *ответы типа "ДА - НЕТ"*; *ответы фиксированно-конструируемого типа* (когда дается часть ответа и необходимо дописать недостающие лексемы); *численные ответы и т.п.*

II. Вопросы, требующие раскрытия в ответе *ТО* одного главного понятия. Это вопросы следующего типа: *'Что выполняется раньше: компиляция или загрузка?'*, *'Назовите предназначение информационного обеспечения САПР'* и т.п.

Выделяются следующие *классы ответов, раскрывающие одноименные ТО: СОСТАВ, ВКЛЮЧЕНИЕ, ДЕЙСТВИЕ, СОСТОЯНИЕ, ВРО, ПРО, КЛО, КЧО*. Выделение классов ответов не зависит ни от предметной области, ни от понятий *ПО*.

СОСТАВ — это класс ответов, в которых понятие раскрывается через его *ТО СОСТАВ* с составляющими понятиями: *' S_i состоит из S_{i+3} и S_{i+4} '*

ВКЛЮЧЕНИЕ — это класс ответов, в которых понятие раскрывается через *ТО ВКЛЮЧЕНИЕ* к другому понятию, содержащему его как часть: *' S_i включается в S_{i+1} '*

СОСТОЯНИЕ — это класс ответов, в которых понятие раскрывается через его *состояние (ТО СОСТОЯНИЕ)*: *' S_i играет'*

ДЕЙСТВИЕ — это класс ответов, в которых понятие раскрывается через отношение действие (ТО ДЕЙСТВИЕ) к другим понятиям: ' S_i перемещает S_{i+2} '

ВРО — это класс ответов, в которых понятие раскрывается через его отношение (ТО ВРО) к другим понятиям по некоторым временной шкале: ' S_i выполняется раньше S_{i+1} '

ПРО — это класс ответов, в которых понятие раскрывается через отношение (ТО ПРО), отражающее его положение в пространстве относительно других понятий: ' S_i находится на S_{i+2} '

КЛО — это класс ответов, в которых раскрывается количественное отношение (ТО КЛО) понятия к другим понятиям: ' S_{i+1} больше, чем S_i '

КЧО — это класс ответов, в которых раскрывается качественное отношение понятия к другим понятиям: ' S_{i+2} легче, чем S_i '

III. Вопросы, требующие раскрытия в ответе СО одного главного понятия. К этому типу относятся вопросы: 'Какую функцию выполняет компилятор?', 'Назовите предназначение загрузчика' и т.п.

Такому типу вопросов соответствуют классы ответов, в которых главное понятие раскрывается через СО.

Например, выделен класс ответов **ФУНКЦИЯ**. В этом классе ответов **ФАКТОР** раскрывается через его СО **ФУНКЦИЯ**: ' S_i переводит S_{i+6} и S_{i+7} '

IV. Вопросы, требующие раскрытия в ответе произвольной комбинации ТО и/или СО одного главного понятия. К данному типу относятся вопросы: 'Что вы знаете о кибернетике?', 'Дайте определение компилятора'. Этим вопросам соответствуют классы ответов, в которых главное понятие раскрывается через его ТО и/или СО. Выделяются следующие классы ответов:

1) **ОПИСАНИЕ** — это класс ответов, в которых раскрываются произвольные комбинации ТО и/или СО главного понятия с другими понятиями: ' S_i состоит из S_{i+3} , S_{i+4} , S_{i+5} . Переводит S_{i+6} и S_{i+7} и выполняется раньше S_{i+1} '.

2) **ОПРЕДЕЛЕНИЕ** — это класс ответов, в которых главное понятие раскрывается через ОП — обобщающее понятие (т.е. понятие на более высоком уровне в иерархии) — и класс ОПИСАНИЕ (т.е. комбинацию ТО и/или СО ОП с другими понятиями). Например, к этому классу можно отнести ответ: 'Студент — это человек, который обучается в вузе'.

3) **ПРИЧИНА** — это класс ответов, в которых раскрывается условие существования некоторых отношений главного понятия с другими понятиями. Предполагается, что главное понятие следствия и его отношения с другими понятиями заданы в вопросе. Например, рассмотрим текст ответа: 'Дерево не тонет в воде, потому что удельный вес дерева меньше удельного веса воды'.

Если это ответ на вопрос: 'Почему дерево не тонет в воде?', то ответ относится к классу ПРИЧИНА. Здесь главное понятие следствия 'дерево' и его отношение с объектом 'вода' дается в самом вопросе. Часть ответа:

'Потому что удельный вес дерева меньше удельного веса воды', — раскрывает условие существования указанного следствия.

4) **СЛЕДСТВИЕ** — это класс ответов, в которых раскрывается следствие от существования некоторых отношений главного понятия с другими понятиями. Предполагается, что главное понятие причины и его отношения с другими понятиями даны в самом вопросе.

Тот же пример в этом случае демонстрирует ответ на вопрос:

'Что следует из того, что удельный вес дерева меньше удельного веса воды?'. Здесь главное понятие причины 'удельный вес дерева' и его отношение 'меньше' к другому понятию 'удельный вес воды' даются в вопросе. В части ответа: 'Дерево не тонет в воде' раскрывается следствие от существования указанного условия.

В ответах на вопросы типа I-IV *главное понятие* не меняется в процессе просмотра текста (т.е. предполагается, что ответы содержат информацию только относительно одного *главного понятия*).

V. Вопросы, требующие раскрытия в ответе более чем одного *главного понятия*. Например, к ним относятся вопросы следующего типа: *'Расскажите о Казанском университете', 'Докажите теорему'* и т.п.

Этому типу вопросов могут соответствовать ответы, в которых *главное понятие* меняется в процессе просмотра ответа. Например, выделены следующие классы ответов, в которых содержатся *главное понятие*, связанные только общим контекстом:

1) **ДЕТАЛИЗАЦИЯ**. В ответах этого класса происходит *детализация понятий, состоящих в некотором отношении с главным понятием*.

Например, пусть задан вопрос V типа: *'Какая связь существует между институтом и заводом?'*. Ответом может быть текст, относящийся к классу **ДЕТАЛИЗАЦИЯ**: *'В институте разработана САПР, которая используется для проектирования токарных приспособлений, которые внедряются на заводе'*. В этом ответе три *главных понятия* — *'институт', 'САПР', 'токарные приспособления'*. Раскрываются следующие отношения: *'институт разработал САПР', 'САПР проектирует токарные приспособления, 'токарные приспособления внедряются на заводе'*.

2) **ПСО (причинно-следственное отношение)**. В ответах данного класса различаются **ФАКТОР** *причинной части ответа, где раскрывается причина, и ФАКТОР части ответа, где раскрывается следствие*. Например, к этому классу относится текст: *'Дерево не тонет в воде, потому что его удельный вес меньше удельного веса воды'* при условии, что в вопросе явно не задано ни одно из отношений. В этом ответе *главное понятие причинной части: 'его удельный вес. Главное понятие части ответа, где раскрывается следствие, есть 'дерево'*. Сердцевинной методологии семантической классификации, предложенной в работе и отличающей ее от традиционных *WH-классификаций*, является разбиение текстов на семантические классы по типу отношения *главного понятия*, раскрываемого в данном ответе. Такая классификация не зависит ни от конкретной ПО, ни от понятий данной ПО, ни от конкретного языка общения с системой. Это позволяет строить

эффективные предметно-независимые анализаторы, ориентированные на раскрытие определенного типа отношения *главного понятия* в рамках соответствующего класса ответов. Также как и в случае выделения типов отношений, каждая группа классов ответов I-V может быть расширена АВТОРОм.

В разделе 3.2.3 раскрывается методология разработки правильных формул передачи ожидаемого смысла ответа на вопрос — индивидуальных концептуальных грамматик, а также конкретные ИКГ. Семантическим классам ответов соответствуют присущие им схемы сочетания (формулы) концептуал, отражающие канонический смысл ответов данного класса (значений вопросов). Канонический смысл ИКГ расшифровывает однозначно. Например, в сочетании концептуал $SS \rightarrow R_{SO} \rightarrow SO \rightarrow SO \rightarrow SO$ передается смысл: *'Главное понятие находится в некотором отношении с тремя другими понятиями. Дано отношение главного понятия - R_{SO} '*. Этот канонический смысл не зависит ни от ПО, ни от понятий, ни от классов ответов, основанных на данной ИКГ. Таким образом, смысл введения ИКГ заключается в сведении семантического анализа текста к синтаксическому анализу его канонического представления в условиях, определенных некоторым контекстом.

Совокупность ИКГ составляет концептуальную грамматику, ориентированную на контекст. В работе в нотации Бэкуса-Наура описываются ИКГ для всех выделенных семантических классов ответов. Например, классы ответов, раскрывающих ТО понятий, имеют следующую ИКГ (обозначим ИКГ2):

$$\begin{aligned} <ИКГ2>:: = SS^* \rightarrow R_{SO}^* \rightarrow SO^* \mid R_{SO}^* \rightarrow SS^* \mid SS^* \rightarrow R_{SO}^* \mid \\ SO^* \rightarrow R_{OS}^* \rightarrow SS^* \mid R_{OS}^* \rightarrow SO^* \mid SO^* \rightarrow R_{OS}^* \mid (R_{SO} \rightarrow Gm)^* \rightarrow SO^* \mid \\ R_{SO}^* \rightarrow (SO \rightarrow Gm)^* \mid (R_{SO} \rightarrow Gm)^* \rightarrow SO^* \mid R_{OS} \rightarrow (SS \rightarrow Gm)^* \end{aligned}$$

Знак '|' обозначает альтернативное вхождение сочетаний концептуал. Круглые скобки (.) служат для объединения концептуал разных типов.

ИКГ классов ответов на вопросы типа III являются определенной (фиксированной) комбинацией ИКГ классов ответов на вопросы типа II. В этих ответах значащими являются понятия и СО между ними.

Рассмотрим возможные ответы на вопрос типа III: *'Какую функцию выполняет компилятор?'*

- 1) переводит исходный текст на языке высокого уровня в объектный текст в машинных кодах,
- 2) получает ЯМК из ЯВУ,
- 3) компилятор переводит ЯВУ в ЯМК.

Здесь отношение 'переводит' есть R_A , отношение 'получает' - R_P , понятия 'текст на языке высокого уровня', 'ЯВУ' - SA , 'текст в машинных кодах', 'ЯМК' - SP , предлог 'из' - GP_A , предлог 'в' - GP_P , понятие 'компилятор' есть главное понятие - SS .

Формализованное представление ответов имеет вид:

- 1) $R_A \rightarrow SA \rightarrow GP_P \rightarrow SP$
- 2) $R_P \rightarrow SP \rightarrow GP_A \rightarrow SA$
- 3) $SS \rightarrow R_A \rightarrow SA \rightarrow GP_P \rightarrow SP$

Исследуя и обобщая формулы таким образом, составляем следующее описание ИКГ классов ответов ФУНКЦИЯ:

$$\begin{aligned} \langle \text{ИКГ ФУНКЦИЯ} \rangle ::= & [SS^* \rightarrow]((R_A^* \rightarrow (GP_P \rightarrow SP^* \rightarrow SA^* | \\ & SA^* \rightarrow GP_P \rightarrow SP^*) | RP^* \rightarrow (GP_A \rightarrow SA^* \rightarrow SP^* | \\ & SP^* \rightarrow GP_A \rightarrow SA^*))((GP_P \rightarrow SP^* \rightarrow R_A^* \rightarrow SA^* | \\ & SA^* \rightarrow R_A^* \rightarrow GP_P \rightarrow SP^*) | (GP_A \rightarrow SA^* \rightarrow R_P^* \\ & \rightarrow SP^* | SP^* \rightarrow R_P^* \rightarrow GP_A \rightarrow SA^*))) \end{aligned}$$

Квадратные скобки [,] означают необязательное вхождение.

ИКГ классов ответов, соответствующих вопросам типа IV-V, являются произвольной комбинацией ИКГ классов ответов на вопросы типа II и III.

Рассмотрим, например, ответ на вопрос IV типа:

'Расскажите о компиляторе'.

Ответ: *'Состоит из лексического и семантического анализатора, включает также этапы генерации и оптимизации кодов. Переводит ЯВУ в ЯМК и предшествует этапу загрузки. Компилятор входит в состав операционной системы и находится в библиотеке системы'.*

В этом ответе класса ОПИСАНИЕ значащими являются понятия: 'лексического', 'синтаксического', 'генерации', 'оптимизации кодов', 'операционной системы', 'библиотеке системы' - SO, 'компилятор' - SS, отношения 'компилятора': отношения 'состоит из' и 'включает' - TO СОСТАВ (R_{SO}), отношение 'переводит' - CO ФУНКЦИЯ (R_A), отношение 'предшествует' - TO ВРО (R_{SO}), отношение 'входит в' - TO ВКЛЮЧЕНИЕ (R_{SO}), отношение 'находится в' - TO ПРО (R_{SO}); понятия 'ЯВУ' - SA, 'ЯМК' - SP, предлог 'в' - GP_P .

Формализованное представление данного ответа имеет следующий вид:

$$[SS \rightarrow] R_{SO} \rightarrow SO \rightarrow SO \rightarrow [SS \rightarrow] R_{SO} \rightarrow SO \rightarrow SO \rightarrow [SS \rightarrow] R_A \rightarrow SA \rightarrow GP_P \rightarrow SP \rightarrow [SS \rightarrow] R_{SO} \rightarrow SO \rightarrow SS \rightarrow R_{SO} \rightarrow SO \rightarrow [SS \rightarrow] R_{SO} \rightarrow SO$$

Таким образом, данное представление есть комбинация шести цепочек концептуал, раскрывающих шесть отношений понятия 'компилятор': СОСТАВ - СОСТАВ - ФУНКЦИЯ - ВРО - ВКЛЮЧЕНИЕ - ПРО соответственно. Обобщая исследования формулы классов ответов на вопросы типа IV, получаем следующее описание ИКГ классов ответов ОПИСАНИЕ, ПРИЧИНА, СЛЕДСТВИЕ (обозначим ИКГ4):

$$\langle \text{ИКГ4} \rangle ::= \langle T2 \rangle * | \langle T3 \rangle * | \langle T2 \rangle * \rightarrow \langle \text{ИКГ4} \rangle * | \langle T3 \rangle * \rightarrow \langle \text{ИКГ4} \rangle$$

Здесь $T2$ обозначает ИКГ классов ответов на вопросы типа II; $T3$ - ИКГ классов ответов на вопросы типа III.

Рассмотрим ответ, относящийся к классу ДЕТАЛИЗАЦИЯ:

' S_i переводит S_{i+1} в S_{i+2} , который находится на S_{i+3} , который выполняется раньше S_{i+4} , который больше, чем S_{i+6} , который стирается'.
(a1)

Формализованное представление ответа имеет следующий вид:

$SS(1) \rightarrow SS(1)DET \rightarrow R_A \rightarrow SA \rightarrow GP_P \rightarrow SP \rightarrow SS(2)DET \rightarrow R_{SO} \rightarrow SO$
 $\rightarrow SS(3)DET \rightarrow R_{SO} \rightarrow SO \rightarrow SS(4)DET \rightarrow R_{SO} \rightarrow Gm \rightarrow SO \rightarrow SS(5)DET$
 $\rightarrow R_{SO}$

Здесь концептула $SS(1)$ есть первое главное понятие - отражает понятие S_i , S_{i+1} есть SA , второе главное понятие $SS(2)=SP$ отражает понятие S_{i+2} , третье главное понятие $SS(3)=SO$ - S_{i+3} , четвертое главное понятие $SS(4)=SO$ - понятие S_{i+4} , пятое главное понятие $SS(5)=SO$ - понятие S_{i+5} . Отношение 'переводит' есть СО ФУНКЦИЯ - R_A ; 'находится на', 'раньше', 'больше', 'стирается' относятся, соответственно, к ТО ПРО, ВРО, КЛО и СОСТАВ. Все $ТО$ обозначаются концептулой R_{SO} . Предлог 'в' есть GP_P .

Таким образом, обобщая формулы ИКГ класса ДЕТАЛИЗАЦИЯ (обозначим ИКГ5) получаем следующее описание:

$\langle \text{ИКГ5} \rangle ::= ([SS(i) * \rightarrow] SS(i)DET \rightarrow \langle G(i) \rangle *)$.

Здесь $SS(i)$ - концептула, отражающая i -е главное понятие; $SS(i)DET$ обозначение начала детализации i -го главного понятия; $G(i)$ - часть ИКГ класса ДЕТАЛИЗАЦИЯ с постоянным $SS(i)$, т.е. грамматика $G(i)$ есть ИКГ4 и раскрывает только i -е главное понятие.

В разделе 3.4 дается описание индивидуальных концептуальных грамматик как основы семантической интерпретации для татарских вопросно-ответных текстов. В рамках той идеологии, которой мы придерживаемся при разработке систем обработки информации, а именно, возможности применения их как в контексте русского, так и татарского языков, среди принципов построения лингвистического процессора в вопросно-ответной ситуации выделены языкозависимые, т.е. требующие их модификации или замены при смене языка. В силу идентичности понятия "здоровый смысл" (т.е. объектов и отношений ПО, возможности их отражения определенными последовательностями определенных смыслообразующих единиц и др.) для русского и татарского языков, идентичными являются все 8 принципов (2 методологических и 6 реализационных), описанные ранее. Отличия обнаруживаются на уровне выделения конкретных концептул, семантических классов ответов, ИКГ. Отличие в наборе концептул вызвано, прежде всего, тем, что в состав концептул включены также грамматические признаки.

Известно, что разные языки используют различный синтаксический строй для передачи смысла предложения. В частности, русский язык относится к языкам типа *SVO* (subject-verb-object), в то время как татарский язык, к языкам типа *SOV* (subject-object-verb). Соответственно, и ИКГ, как грамматики 'смыслообразующих единиц' (т.е. концептул), отражают специфику языков и также различаются для рассматриваемых нами русского и татарского языков. Например, исследования ИКГ для классов ответов на вопросы типа III на татарском языке приводят к следующему обобщенному описанию:

ИКГ(Т) ФУНКЦИЯ :: $SA \rightarrow R_A \rightarrow SP \rightarrow R_P | SP \rightarrow R_P \rightarrow SA \rightarrow R_A |$
 $SA + Gm \rightarrow SP \rightarrow R_P | SP \rightarrow SA + Gm \rightarrow R_P | SP \rightarrow SA \rightarrow R_A \rightarrow R_P | SP \rightarrow R_P$
 $\rightarrow SP \rightarrow SA \rightarrow R_A \rightarrow R_P | SP \rightarrow R_P \rightarrow SP \rightarrow SA + Gm \rightarrow R_P | SA \rightarrow SP \rightarrow R_P$
 $\rightarrow R_A | SP \rightarrow R_P$

Введение правила корректного преобразования (не изменяющего “глубинного смысла”) $SA + Gm = SA \rightarrow R_p$ позволяет оптимизировать грамматику следующим образом: ИКГ(Т) ФУНКЦИЯ::= $SA \rightarrow R_A \rightarrow SP \rightarrow R_p \mid SP \rightarrow R_p \rightarrow SA \rightarrow R_A \mid SP \rightarrow SA \rightarrow R_A \rightarrow R_p \mid SP \rightarrow R_p \rightarrow SP \rightarrow SA \rightarrow R_A \rightarrow R_p \mid SA \rightarrow SP \rightarrow R_p \rightarrow R_A \mid SP \rightarrow R_p$

В главе 4 описывается Татарский морфологический анализатор. Здесь дается описание двухуровневой модели морфологии татарского языка и морфологического анализатора, построенного на ее основе с использованием программного инструментария PC-KIMMO и относящегося к классу прагматически-ориентированной концептуально-формальной модели. В разделах 4.2 и 4.3 дается подробное описание файлов правил генерации и морфотактических правил для распознавания словоформ татарского языка с демонстрацией материала на многочисленных примерах. Для описания файла фонологических правил татарского языка используется 39 записи правил. Приведем пример описания соответствия лексического символа *A* поверхностным символам *a* и *ə*, отражающего закон сингармонизма в татарском языке. Правило (1) для лексического символа *A* имеет следующий вид:

(1) ПРАВИЛО $A:a \Rightarrow \text{VOWEL:BACKV CS:@* @:0* +:0* +:й* +:н* _ | BACKV CS * +:0* CS* _ | W:a +:0 Л:л _ | [W:й|W:a] +:0* [Л:л|ч:ч] _ | BACKV +:0* CS* +:0* K:@ _ | `:0 CS:@* +:0 @:0* CS* _}$

Символ * означает возможность появления указанного соответствия 0 и более раз. Символ ‘|’ в (1), обозначающий логическое ИЛИ, используется в правиле для разделения вариантов проявлений одного и того же соответствия в разных контекстах. Такая многовариантность контекстов вызвана либо нарушением морфологии татарского языка (что требует ввода дополнительных специальных символов и правил), либо существованием лексических символов, имеющих собственные правила описания, и которые невозможно заранее определить в множествах BACKV (гласные буквы заднего ряда) и FRONTV (гласные буквы переднего ряда).

Таким образом, рассматриваемое Правило (1) включает 6 левых контекстов (т.е. зависит только от предыдущих символов):

- (а) $\text{VOWEL:BACKV CS:@* @:0* +:0* +:й* +:н* _}$;
- (б) $\text{BACKV CS * +:0* CS* _}$;
- (в) W:a +:0 Л:л _ ;
- (г) $\text{[W:й|W:a] +:0* [Л:л|ч:ч] _}$;
- (д) $\text{BACKV +:0* CS* +:0* K:@ _}$;
- (е) $\text{`:0 CS:@* +:0 @:0* CS* _}$.

Здесь CS - множество согласных букв, @ - любой символ алфавита. Согласно определению фонологических правил, Правилom (1) в рамках контекста (а) утверждается следующее: лексический символ *A* соответствует поверхностному *a*, если ему предшествует любая гласная из множества заднего ряда гласных BACKV; далее может следовать согласная буква из множества CS, которая соответствует любому символу из татарского алфавита. Далее идет описание контекста для случаев, когда в татарском

языке при присоединении аффиксов символ + либо просто аннулируется, либо используется как дополнительный символ.

Рассмотрим примеры с контекстом (а). Пусть на вход генератора подаются следующие последовательности морфем:

бала+ЛАР 'ребенок' + [аффикс множественности]

бала+ДА 'ребенок' + [аффикс место-временного падежа (локатив)]

Происходит следующая обработка входной последовательности. Вначале генератор устанавливает промежуточные формы *балалАр* и *баладА*, соответственно, которые получаются благодаря определенному заранее списку предполагаемых пар (т.е. б:б, а:а, л:л) и отдельным правилам для лексических символов Л и Д, имеющих соответствие Л:л и Д:д в этом контексте.

Левый контекст символа А в обоих случаях удовлетворяет условиям контекста (а) Правила (1), т.е. символу А предшествует согласная буква, слева от которой следует гласная буква заднего ряда. Следовательно, генератор выдает следующие поверхностные формы записи:

балалар 'дети'

балада 'у ребенка'

Файл морфотактических правил разработан на основе морфотактических схем и определяет взаимосвязи между основной и аффиксальными группами. Лексикон корневых лексем построен на основе современного татарского языка и состоит из ряда лексиконов, заполненных согласно соответствующих требований РС-KIMMO. Подлексиконы содержат строки лексических входов, состоящих из следующих трех частей: первая часть - лексический атом (татарское корневое слово); вторая часть - класс присоединения (или продолжения); т.е. то, что может присоединяться, следовать как продолжение непосредственно за этим атомом - подлексикон, который может иметь другие лексические единицы. Классы присоединений могут следовать за множеством других морфемных единиц. Лексикон ALTERNATION в РС-KIMMO — это список названий подлексиконов, порядок которых определяет — какой класс за каким может следовать, притом, возможно только одно его определение, то есть это ограничение, свойственное подлексикону; третья часть - его трактовка (описание грамматических признаков). Как правило, здесь записываются любые морфологические, грамматические, или семантические свойства лексической единицы. При обработке слова распознавателем трактовка каждой избранной морфемы добавляется в строку результата. В диссертации определен и описывается следующий список лексиконов для татарского языка.

(1) Имена существительные (Nouns). Лексикон включает около 15500 корневых имен существительных.

(2) Глаголы (Verbs). Лексикон содержит около 7100 глагольных корней.

(3) Прилагательные (Adjectives). Как известно, татарский язык является агглютинативным регулярным языком, подчиняющимся строгим правилам. Вместе с тем, как и в любом естественном языке, имеются исключения, чаще всего, также подчиняющиеся определенным закономерностям. Так, прилагательные превосходной степени имеют префиксы, записываемые через

дефис '-'. Например: корневое слово *кызыл* ('красный') в превосходной степени записывается как *кып-кызыл* ('очень красный'). Лексикон Прилагательные содержит около 3150 базовых корней и дополнительно включает лексикон, состоящий из 138 прилагательных превосходной степени с префиксами. Определены также следующие Лексиконы, составляющие небольшую долю в общем словаре, включающем 25 900 корневых слов, имеющие особые морфотактические правила, присущие выделенным группам слов: (4) Наречия (Adverbs). (5) Местоимения (Pronouns). (6) Числительные (Numerals). (7) Послелogi (Postpositions). (8) Союзы (Conjunctions). (9) Междометия (Exclamations).

Параметр ALTERNATION включает 8 входов для словоформ (т.е. в данном описании определено, что имеет место 8 разных возможностей для начала татарского слова): VERB (Глагол) - подлексикон для глаголов; NOUN (Имя существительное) - подлексикон для существительных; ADJECTIVE (прилагательное) - подлексикон для прилагательных; ADJECTIVE2 (прилагательное2) - подлексикон для прилагательных; NUMERAL (число) - подлексикон для чисел; PRONOUN (местоимение) - подлексикон для местоимений, послелогов; ADVERB (наречие) - подлексикон для наречий; SPECIAL (специальное) - подлексикон для союзов, междометий. Дается подробное описание базы морфотактических правил для всех типов корневых слов с примерами и комментариями. Список слов для распознавания записывается в специальный файл <tatar.rg>, который подается на вход распознавателя двухуровневого морфологического анализатора. Пусть, например, файл <tatar.rg> содержит следующие глагольные формы: *baru bargannar barma barmasa bardirt*

В файле <tatar.rec> получаются следующие результаты распознавания:

bar +V	[V(go)+NOMINATIVE(u/U)]
bar +KEн+LEr	[V(go)+PAST_UNDEF(KEн)+PLURAL(LEr)]
bar +mE	[V(go)+NEGATIVE(mE)]
bar +mE+sE	[V(go)+NEGATIVE(mE)+CONDITIONAL(sE)]
bar +DHr+t	[V(go)+COUSATIVE(DHr)+MODAL(t)]

В выводах подчеркивается, что двухуровневая модель морфологии татарского языка относится к классу прагматически-ориентированных концептуально-формальных моделей и представляет собой полную компьютерную модель татарской морфологии.

Татарский морфологический анализатор, созданный с использованием программной оболочки PC-KIMMO, может быть реализован в составе других систем обработки ЕЯ-текстов. В частности, в настоящее время создана рабочая версия программы татарско-турецкого машинного переводчика на основе татарского и турецкого морфологического анализаторов, татарско-турецкого электронного словаря и сервисных программ. Морфологический анализатор на базе PC-KIMMO может эффективно использоваться также и как программный инструмент изучения и развития морфологии татарского языка.

Глава 5 содержит описание морфологического корректора татарских текстов ТАТКОР, первого корректора, осуществляющего автоматизированную

проверку правильности написания татарских словоформ на основе генеративной морфологии, представленной в виде диаграмм словоизменений по частям речи, и орфографического словаря татарского языка. ТАТКОР обнаруживает ошибки в текстовых файлах и позволяет исправлять их как в режиме редактора, так и с помощью списка альтернатив, сформированного программой.

Эффективность морфологического корректора обеспечивается следующими его характеристиками: меню-диалоговое общение с конечным пользователем; специальная организация словаря с указанием признака принадлежности к определенной части речи (словарных статей); формирование всевозможных правильных словоформ на основе специальных оригинальных диаграмм - схем сочетания аффиксов и сопоставление их с заданной словоформой; выделение некорректных словоформ другим цветом (красным); высвечивание на экране информации для самоконтроля: количество проверенных слов, количество ошибочных слов в тексте, количество слов, занесенных в персональный словарь; наличие персонального словаря пользователя, который позволяет расширять базовый словарь необходимой лексикой; наличие режима подсказки, предоставляющей возможные правильные варианты проверяемой словоформы по запросу пользователя; возможность проверки текстовых файлов практически неограниченной длины; возможность инициализации проверки текста с любой указанной строки; оптимальная организация словарной базы, включающей три словаря основ: персональный и внутренний словари, располагающиеся в оперативной памяти, и внешний словарь - на жестком диске.

Существующие корректоры русских текстов, как правило, в качестве словарной базы используют специальные парадигматические словари, разрабатываемые на основе словаря Зализняка. Применение такого подхода вполне оправдан в случае проверки корректности словоформ флексивного типа языков, в которых каждая основа или корень имеют ограниченное и конечное количество словоизменений и их можно описать с помощью соответствующих парадигм, с указанием всех возможных вариантов. Другая ситуация для языков агглютинативного типа, в которых количество присоединяемых аффиксов потенциально не ограничено и практически невозможно исчерпывающе описать словоформы в виде основ и соответствующих парадигм. При анализе татарских текстов установлено, что в 90% случаев к слову присоединяется не более 3 аффиксов, но в некоторых случаях возможно участие гораздо большего числа аффиксов (более 10) в образовании татарской словоформы. Число порождаемых словоформ по одной основе для татарского языка более 50 тысяч (формально не ограничено), в то время как для русского примерно несколько десятков, а для английского языка это соотношение составляет 1,6. Морфология татарского языка, к тому же, однозначно определяет место каждого типа аффиксов в словоформе.

В связи с этим для создания корректора татарского языка выбран генеративный подход к формированию словоформ заключается в следующем. Из списка основ, составленного после поиска по словарю, отбирается основа,

имеющая наибольшую длину, и соответствующая процедура пытается присоединить к ней аффиксы из некоторого списка. Аффикс присоединяется, если основа с присоединенным аффиксом имеет вхождение в проверяемое слово. Таким образом, получается новая словоформа с измененным типом, и к нему могут присоединиться аффиксы из списка, соответствующего новому типу. Это действие продолжается до тех пор, пока не получится проверяемое слово, или список аффиксов не окажется исчерпанным. Если проверяемого слова не получилось, то берется очередная основа из списка основ и с ней повторяются те же действия. При обнаружении неправильных слов корректор "подсказывает" вероятные корректные варианты, в которых, возможно, пользователем допущена ошибка. По неправильному слову с помощью операций *вставка, замена, удаление буквы и обмен соседних букв* генерируются слова, которые подвергаются проверке. Правильные кандидаты предлагаются в качестве подсказки. Особенности татарского языка и явление рекурсии, что является главным препятствием к применению в корректоре парадигматического подхода, подробно рассмотрены в главе 2.

Пример. Рассмотрим корневое слово: *урман* - 'лес'. Следующие словоформы, образованные из данного слова присоединением аффиксов - *ныкы* и -*дагы* потенциально являются неограниченными: *урман-дагы-ныкы-ндагы-ныкы-ныкы-(дагы)-(ныкы)*. Такая словоформа имеет определенный смысл, который на русском языке будет передаваться длинным рядом словоформ. Приведем несколько иллюстраций: *урман* - 'лес', *урмандагы* - 'то, что/кто в лесу', *урмандагыныкы* - 'нечто того, что/кто в лесу', *урмандагыныкындагы* - 'на том, что нечто того, что в лесу', *урмандагыныкындагыныкыныкы* - 'нечто то, того, что нечто то, на том, что нечто того, что в лесу'.

Аффиксы в скобках обозначают продолжающийся ряд. Причем, между этими аффиксами могут появляться другие аффиксы, скажем, аффикс множественности -*лар*. Например, *урмандагыларныкыларныкы* - 'то нечто, тех нечто, что в лесу'.

Разработка корректора на основе генеративной модели оправдана также следующими дополнительными возможностями:

- вместе с основной словоформы из словаря извлекаются также и дополнительные характеристики, существенно упрощающие процесс коррекции, т.е. позволяющие существенно сократить пространство возможных комбинаций аффиксов;

- обеспечивается эффективное развитие корректора как программы синтаксической проверки, при котором является очевидным избыточность полного перебора всех возможных сочетаний аффиксов. Формы синтагм во многом будут определяться сочетаемостью соответствующих лексем.

ЗАКЛЮЧЕНИЕ

Диссертационная работа посвящена проблеме разработки лингвистических моделей и созданию на их основе систем и информационных технологий автоматизированной обработки ЕЯ-текстов.

Главными научными результатами диссертации являются: 1) разработка технологии построения прагматически-ориентированных лингвистических моделей и 2) построение компьютерной модели морфологии татарского языка.

Первый результат позволяет утверждать, что получено новое решение актуальной проблемы повышения эффективности систем и технологий обработки ЕЯ-текстов за счет разработки лингвистических моделей, включающих изначально минимальный набор языковых средств и определяющих детальность их описания исходя из методологических принципов “детерминированности контекста” и “ожидаемости ответа”. Технология построения прагматически-ориентированных лингвистических моделей определяет объект приложения, необходимый инструментарий, последовательность и содержание этапов создания лингвопроцессоров на основе соответствующих моделей.

Второй результат позволяет утверждать, что впервые получено решение важной и актуальной задачи создания компьютерной модели морфологии татарского языка как формальной базы лингвопроцессоров различного назначения.

Основным практическим результатом диссертации является разработка прагматически-ориентированных лингвистических моделей, их программная реализация в виде систем обработки информации и широкое их внедрение для решения задач автоматизированной обработки ЕЯ-текстов.

Кроме того, в диссертации описаны следующие основные результаты работы:

1. Сформулирована проблема, разработана методология и предложены методы и базовые принципы и технология построения систем обработки текстовой информации, основанных на прагматически-ориентированных моделях, или моделях “от цели”.

2. Разработана и реализована структурно-функциональная модель морфем татарского языка, относящаяся к классу концептуально-функциональных моделей, и представляющая собой программно-информационный инструментарий, используемый как для построения прагматически-ориентированных лингвопроцессоров, так и в качестве информационно-справочной и учебно-консультационной базы.

3. На основе анализа и обобщения средств формального представления семантики контекста, необходимого для изучения и описания значений аффиксальных морфем, проведена классификация и описаны в виде семантических сценариев базовые типы соответствующих контекстов. Также разработана объектно-предикатная система, включающая базовые отношения, используемые для описания значений аффиксальных морфем в структурно-функциональной модели.

4. Разработана открытая прагматически-ориентированная модель анализа ЕЯ-текстов вопросно-ответном контексте, относящаяся к диалоговым моделям, и реализован двухуровневый семантический интерпретатор. Эффективная реализация модели достигнута за счет следующих результатов: - выделена система смыслообразующих единиц (концептул), отражающих типовые понятия, базовые отношения, некоторые грамматические признаки и специальные лексемы, с целью трансформации проблемы семантического анализа вопросно-ответного текста в проблему синтаксического анализа в условиях детерминирующей роли контекста; - описаны методология семантической классификации вопросно-ответных текстов на основе типовых отношений, а также базовые типы вопросов и классы ответов; - разработана методология и приведено описание индивидуальных концептуальных грамматик, представляющих собой логически правильные смысловые конструкции ответов соответствующих классов для татарского и русского языков.

5. Разработана двухуровневая модель морфологии татарского языка, относящаяся к классу прагматически-ориентированных концептуально-формальных моделей, включающая фонологические и морфотактические правила, и представляющая собой полную компьютерную модель татарской морфологии. На ее основе построен двухуровневый морфологический анализатор, который осуществляет генерацию и разложение произвольных словоформ на морфемы.

6. Разработана и реализована генеративная модель морфологии татарского языка как основа морфологического корректора, позволяющая учитывать такие особенности татарского языка, как агглютинативность, наличие рекурсии.

В десяти приложениях приведены описания конкретных примеров реализации по каждой разработанной прагматически-ориентированной модели, справки и акты о внедрении и применении результатов диссертационной работы.

Основные публикации по теме диссертационной работы:

1. Сулейманов Д.Ш. О семантическом анализе текстов в АОС // Современные ЭВМ. Разработка, производство и технологические процессы изготовления ЭВМ. Использование ЭВМ. Мат.обеспечение ЭВМ и АСУ: Тез. докл. 5 научно-техн. конференции молодых ученых и специалистов. - Казань: Татполиграф. комб. им. К.Якуба, 1980. - С.46-47.
2. Сулейманов Д.Ш. Об одном подходе к семантическому анализу текстов в АОС // Тез. докл. конф. молодых ученых по вопросам радиоспектроскопии; квантовой акустики, механики и прикладной математики. -Казань: Изд-во Казан. физ.-тех. ин-та КФАН СССР, 1980. -С.113-114.
3. Бухараев Р.Г., Сулейманов Д.Ш. К разработке АОС с развитыми возможностями семантического анализа ответов обучаемого //АОС и их применение: Тез. докл. межвуз. совещания-семинара "Проблемы

- автоматизации обучения персонала систем организационного управления". - Казань: Изд-во Казан. ун-та, 1983. - С.7-9.
4. Сулейманов Д.Ш. О семантическом анализе текстов в АОС //Вероятностные методы и кибернетика. -Вып. 20. - Казань: Изд-во Казан. ун-та, 1984. - С. 106-116.
 5. Сулейманов Д.Ш. Принципы семантической классификации текстов и их анализ по классам // Математические основы и программное обеспечение автоматизации интеллектуальной деятельности. Матер. третьей научн. конф. молодых ученых и специалистов фак. выч. мат. и киберн. /Казанский ун-т. - Казань: -1984. - С.53-56 [Сб. деп. в ВИНТИ N 331-84].
 6. Сулейманов Д.Ш. Об одной реализации программы анализа ответов обучаемого по ключевым параметрам // Математические основы и программное обеспечение автоматизации интеллектуальной деятельности. - Казань, 1984. - С. 29-32 [Сб. деп. в ВИНТИ N 693-85].
 7. Сулейманов Д.Ш. ЛГ - лингвистический процессор ответов обучаемого в АОС // Математические модели и вычислительная техника в управлении учебным процессом высшей школы (Рига, 19-23 мая 1986г.): Тез. докл. и сообщений межвуз. науч.-метод. семинара. - Рига: Изд-во Риж.ПИ, 1986. -С.37.
 8. Бухараев Р.Г., Сулейманов Д.Ш. Подсистема обеспечения интеллектуального взаимодействия в диалоговой системе // Тез. докл. IV Всесоюзной конференции "Диалог Человек-ЭВМ". Ч.2. ИК им. В.М.Глушкова АН УССР (Киев, 8-12 октября 1985г.). - Киев, 1985. - С.11-13.
 9. Бухараев Р.Г., Сулейманов Д.Ш. Об одном подходе к разработке интеллектуальных АОС // Кибернетика, 1986, N 3. - С.42-49.
 10. Ситницкий Э.А., Трояновский С.В., Исмагилов Н.А., Бухараев Р.Г., Сулейманов Д.Ш. Автоматизированная обучающая система ВУЗ-ОСКАР. Описание языка // Пакет прикладных программ ДПЦ 00002-013501-2.-Таллинн.- 1987.-81 с.
 11. Бухараев Р.Г., Моисеев В.С., Сулейманов Д.Ш. АОС как обслуживающая подсистема развивающейся САПР // Казань, 1987.-12с. [Деп. в ВИНТИ N 3453].
 12. Абросимов А., Гатауллин Г.С., Исмагилов Н.А., Сулейманов Д.Ш. и др. Технологический комплекс обучения КГУ: программно-методическое обеспечение учебного процесса // Казань, 1987. - 37 с. [Деп.в НИИ Проблем ВШ N 1324-87].
 13. Абросимов А., Гатауллин Г.С., Сулейманов Д.Ш. и др. Технологический комплекс обучения КГУ: методология обучения с использованием ЭВМ // Казань, 1987. - 43 стр. [Деп. в ВИНТИ N 916-87].
 14. Сулейманов Д.Ш. Индивидуальные концептуальные грамматики //Вероятностные методы и кибернетика. -Вып. 22. -Казань: Изд-во Казан. ун-та. -1987. - С. 105-111.
 15. Бухараев Р.Г., Сулейманов Д.Ш. К разработке семантического анализатора вопросно-ответных текстов // Тез. докл. Всесоюзной школы-семинара. (Москва, 1-2 мая 1990г.). -М., 1990.-С.19-20.

16. Бухараев Р.Г., Сулейманов Д.Ш. Семантический анализ в вопросно-ответных системах. - Казань: Изд-во Казан. ун-та. - 1990. -124 с.
17. Сулейманов Д.Ш., Шафигуллин Р., Урусов И. Об одном эффективном приложении морфологического корректора татарских текстов // Международный семинар CALL Компьютеризация обучения языкам: Тез. докл. (Казань, 20-24 сент., 1993г.). - Казань, 1993. -С.43-44.
18. Сулейманов Д.Ш. Регулярность морфологии татарского языка и типы нарушений в языке // Серия: Интеллект. Язык. Компьютер. -Вып.1. -Казань: Изд-во Казан. ун-та, 1994. -С.77-106.
19. Сулейманов Д.Ш., Шафигуллин Р.Н. Морфологический корректор татарских текстов ТАТКОР // Серия: Интеллект. Язык. Компьютер.-Вып.2. - Казань: Изд-во Казан. ун-та, 1995. -С.86-89.
20. Сулейманов Д.Ш., Курбский Д.Н. Интегрированная среда ЛИНГВИСТ // Татарский язык и новые информационные технологии. Серия: Интеллект. Язык. Компьютер. -Вып.2. -Казань: Изд-во Казан. ун-та. - 1995. -С.94-96.
21. Бухараев Р.Г., Сулейманов Д.Ш. Механизмы ЕЯ-интерфейса в управляемом контексте // Труды Межд. семинара ДИАЛОГ-95: компьютерная лингвистика и ее приложения (Казань, 31 мая - 4 июня, 1995г.). -Казань, 1995. - С.74-87.
22. Бухараев Р.Г., Сафиуллина Ф.С., Галиуллин К.Р., Еникеев А.И., Сулейманов Д.Ш. К концепции Машинного Фонда Республики Татарстан // Татарский язык и новые информационные технологии. Серия: Интеллект. Язык. Компьютер. -Вып.2. -Казань: Изд-во Казан. ун-та, 1995. -С.20-35.
23. Сулейманов Д.Ш., Шафигуллин Р.Н. Морфологический корректор татарских текстов - ТАТКОР // Труды межд. семинара ДИАЛОГ-96: компьютерная лингвистика и ее приложения (Пушино, 4-9 мая, 1996г.). -М., 1996. -С.253-255.
24. Сулейманов Д.Ш. К вопросу о числе татарских падежей // Исследования в компьютерной лингвистике. Серия: Интеллект. Язык. Компьютер. -Вып.3. - Казань: Изд-во "Фэн". -1996. -С.70-84.
25. Сулейманов Д.Ш., Гатиатуллин А.Р. Модель татарской аффиксальной морфемы и ее реализация // Серия: Интеллект. Язык. Компьютер. -Вып.4. - Казань: Изд-во "Фэн". -1996. -113-127.
26. Сулейманов Д.Ш., Галиев И.Х., Хабибуллин А.Р. Алгоритмический подход к обучению правилам словоизменения татарских существительных // Труды Межд.семинара "Искусственный интеллект в образовании", ч.2 (Казань, 1-4 октября 1996г.). -Казань, 1996. -С.173-175.
27. Сулейманов Д.Ш., Бухараев Р.Г. Принципы построения лингвопроцессоров в вопросно-ответном контексте // Научн.труды нац.конференции с межд.участием "Искусственный интеллект-96", т.1 (Казань, 5-11 октября 1996г.). -Казань, 1996. -С.149-155.
28. Suleymanov D.Sh. Towards a definition of Tatar cases via the structural-functional morphemic model // Proceedings of the 'LP'96 Conference on Languages Typology (Prague, 22-24 August, 1996). Prague, 1996. -P. 388-395.

29. Сулейманов Д.Ш., Гатиатуллин А.Р. Функционально-структурная модель татарских морфем как база для лингвопроцессоров // Труды Межд. семинара Диалог-97: компьютерная лингвистика и ее приложения. (Ясная Поляна, Тула, 10-15 июня 1997г.). - М., 1997. -С.266-271.
30. Suleymanov D.Sh. The semantic analyzer as a part of the embedding Teacher's model in Intelligent Tutoring Systems //Proceedins of the Workshop: Embedding User Models in Intelligent Applications. Sixth International Conference on User Modeling (UM97) (Chia Laguna, Sardinia, Italy, 1-5 June, 1997). Chia Laguna, 1997. - P. 48-53.
31. Suleymanov D.Sh., Guilmullin R.A., Guilmy A.A. Two-level phonological rules of Tatar morphology //KDS-97: Шестая международная конференция "Знания-Диалог-Решение": Сб. Научн. трудов в двух томах. Т1. (Ялта, 15-20 сентября 1997г.). – Ялта, 1997. -С.299-305.
32. Сулейманов Д.Ш., Гильмуллин А.А., Гильмуллин Р.А. Двухуровневое описание морфологии татарского языка // Тез. Межд. научн. конф., посвященной 200-летию университета: "Языковая семантика и образ мира" (Казань, 7-10 октября 1997г.). -Книга 2. –Казань: Изд-во Казан. ун-та, 1997. -С.65-67.
33. Сулейманов Д.Ш. Формальная элегантность и естественная сложность морфологии татарского языка // Электронная конф.: Информационные технологии в гуманитарных науках (Казань, 25-31 мая 1998г.). –Казань, 1998. - [HTTP://www.kcn.ru/_tat_ru/universitet/gum_konf/ot7.htm](http://www.kcn.ru/_tat_ru/universitet/gum_konf/ot7.htm).
34. Сулейманов Д.Ш., Гильмуллин А.А., Гильмуллин Р.А. Файл фонологических правил татарского языка //Электрон. конф.: Информационные технологии в гуманитарных науках (Казань, 25-31 мая 1998г.).–Казань,1998. - [HTTP://www.kcn.ru/_tat_ru/universitet/gum_konf/ot7.htm](http://www.kcn.ru/_tat_ru/universitet/gum_konf/ot7.htm).
35. Сулейманов Д.Ш., Гильмуллин А.А., Гильмуллин Р.А. База морфотактических правил для татарского глагола как основа двухуровневого морфологического анализатора // Сб. трудов Международного семинара ДИАЛОГ-98 (Казань, 1-2 июня 1998г.). -Казань, 1998. -С.597-609.
36. Сулейманов Д.Ш., Гатиатуллин А.Р. Интегрированный программно-информационный комплекс "Морфема" // РАИИ. Научный совет по проблеме "Искусственный Интеллект" РАН: В сб. Трудов шестой национальной конференции с международным участием КИИ-98 в трех томах. Т.1.-Пушино, 1998. -С.208-214.
37. Сулейманов Д.Ш., Гатиатуллин А.Р. Формальное описание значений аффиксальных морфем // В Сб. трудов Международного семинара ДИАЛОГ-98 (Казань, 1-2 июня 1998г.). -Казань, 1998. -С.713-725.
38. Сулейманов Д.Ш., Хадиев Р.М., Якушев Р.С. Основы информатики и издательское дело. -Казань: Изд-во Казан. ун-та. - 1998. -78 с.
39. Сулейманов Д.Ш. Обработка ЕЯ-текстов на основе прагматически-ориентированных лингвистических моделей // Сб.под ред. Соловьева В.Д.: Обработка текста и когнитивные технологии. -Вып.3. Труды научного семинара "Когнитивное моделирование" (Пушино, октябрь 1998 г.). -С.205-212.

40. Сулейманов Д.Ш., Гатиатуллин А.Р. К разработке Лемматора татарских словоформ // В сб. Трудов Международного семинара Диалог-99 по компьютерной лингвистике и ее приложениям в двух томах (Таруса, 31 мая -4 июня 1999г.). Т.2. –Таруса, 1999. -С.306-314.
41. Сулейманов Д.Ш. Аналитический обзор отечественных и зарубежных работ в области обработки естественного языка в аспекте прагматически-ориентированного подхода // В электрон. журнале Казанского госуниверситета "Информационные технологии". –Казань,1999. –HTTP: [//www.kcn.ru/tat_en/science/fttc/vol000/st.doc](http://www.kcn.ru/tat_en/science/fttc/vol000/st.doc) (оглавление: contents.htm).
42. Suleymanov D.S. Natural possibilities of the Tatar morphology as a formal base of the NLP. // In Proceedings of the First International Workshop "Computerisation of Natural Languages" (Varna, Sept. 3-7, 1999). –Sofia (Bulgaria): Information Services Plc, 1999. -P.113-117.

Сдано в набор 1.02.00 г. Подписано в печать 1.02.00 г.
Форм.бум. 60 х 84 1/16. Печ.л. 2,75. Тираж 100. Заказ 37.

. Лаборатория оперативной полиграфии КГУ
420045 Казань, Кр. Позиция, 2а

152