АКАДЕМИЯ НАУК РЕСПУБЛИКИ ТАТАРСТАН
Институт прикладной семиотики АН РТ
Российская ассоциация искусственного интеллекта

ФОРМАЛЬНЫЕ МОДЕЛИ И СИСТЕМЫ В ВЫЧИСЛИТЕЛЬНОЙ ЛИНГВИСТИКЕ

Под редакцией
доктора технических наук, профессора
П. И. Соснина,
кандидата технических наук, доцента
О. А. Невзоровой

Казань – 2016

УДК 004.8+81'322
ББК 81.1

Коллектив авторов:
Д. Ш. Сулейманов, О. А. Невзорова, П. И. Соснин, Л. Н. Беляева,
Н. В. Лукашевич, С. Г. Татевосов

Рецензенты:
доктор технических наук, профессор *А. Я. Фридман*
доктор филологических наук, доцент *Г. В. Колпакова*

Формальные модели и системы в вычислительной лингвистике [Текст] / Д. Ш. Сулейманов, О. А. Невзорова, П. И. Соснин, Л. Н. Беляева, Н.В. Лукашевич, С. Г. Татевосов] : Научное издание / Под редакцией П. И. Соснина, О. А. Невзоровой – Академия наук РТ, Институт прикладной семиотики АН РТ. – Казань: 2016. – 187 с. : ил.

ISBN

Монография посвящена различным формальным моделям и системам в вычислительной лингвистике. Представлены обзорные и оригинальные работы, затрагивающие теоретические и прикладные проблемы вычислительной лингвистики, прежде всего вычислительной семантики, а также создания приложений, в которых применяются семантические языковые модели. К числу обсуждаемых вопросов относятся разработка прагматически ориентированных лингвистических моделей и онтолого-лингвистических систем, онтологическое моделирование, разработка лексикографических ресурсов переводчика, моделирование тональности текстов и семантические исследования средств выражения структуры событий в разных языках.

Анализ проблем, затронутых в монографии, показывает широчайший спектр крайне интересных задач и приложений для новых исследователей, и авторы надеются, что работа внесет свой посильный вклад в вычислительную лингвистику, в которой решения многих задач, связанных с моделированием семантики, еще только формируются.

Для научных работников, преподавателей, аспирантов и студентов, специализирующихся в области вычислительной лингвистики и ее приложений.

УДК 004.8+81'322
ББК 81.1

ISBN

ПРЕДИСЛОВИЕ

Не так далёк тот день, когда расширяющаяся компьютеризация всех сфер человеческой активности, включая различные виды персональной и коллективной деятельности, приведёт к образованию социо-киберфизической ноосферы, в которой будет жить всё население Земли. В образовании такой ноосферы принципиальна роль прогресса телекоммуникаций, который за последнее десятилетие, освоив различные формы связывания людей, включая социальные сети (Internet-of-People), перешёл на то что их окружает как вещи (Internet-of-Things) и как сервисы (Internet-of-Services).

С развитием разновидностей систем расширенного Интернета, уже осуществляется переход к всеохватывающему Интернету (Internet-of-Everythings), а от него в Кибер-сети, где различные неодушевленные предметы и физическая реальность, за счет использования разнообразных вычислительных ресурсов и сенсорных возможностей, наделяются подобием интеллектуальных функций и поведением.

Всё это приводит к новым постановкам вопросов о формах существования людей в глобально информатизируемом мире и, в первую очередь, об их взаимодействии с их окружением. Следует отметить, что в таких взаимодействиях следует сохранять естественность отношений людей с их окружением, естественность в том смысле, что применения кибер-посредников должны включаться в жизнь человека согласовано с её природной интеллектуальной сущностью. А эта сущность определяется всеохватывающем применением естественного языка в жизни по образцам интеллектуально обработанных условных рефлексов, то есть в формах жизни, использующих человеческий опыт.

Именно по этой причине, в последнее время, в среде учёных и практиков наблюдается активный интерес к естественному языку и расширяющийся фронт междисциплинарных исследований и разработок по тематике NLP (Natural Language Processing), результаты которых обсуждаются на многочисленных конференциях и публикуются в периодических журналах. К такому классу публикаций относится и настоящая монография, включающая обзорно-аналитические материалы и тематические публикации по указанному направлению.

Сложность задач лингвистического анализа текстов связана с неодинаковой разработанностью алгоритмического обеспечения различных этапов лингвистического анализа. Разработка алгоритмического обеспечения опирается на формальные модели, при этом разные этапы лингвистического анализа текстов формализованы в разной степени. Наиболее формализованным является морфологический уровень, который описывается различными формализмами (модели конечных автоматов, регулярные грамматики и др.). Несомненно, более сложной является задача моделирования семантики. Язык является сложнейшим объектом для семантического моделирования. Назовем лишь некоторые из наиболее фундаментальных свойств естественного языка: принципиальная нечеткость значения языковых выражений; динамичность языковой системы; образность номинаций, основанная, прежде всего, на метафоричности; креативность в освоении новых знаний; семантическая мощь словаря, позволяющая выражать любую информацию с помощью конечного множества элементов; гибкость в передаче информации; разнообразие функций; специфическая системность.

Можно выделить ряд направлений, в которых развиваются формальные семантические теории: лексическая семантика, формальная семантика, вычислительная семантика, когнитивная семантика, онтологическая семантика.

Монография посвящена различным формальным моделям и системам в вычислительной лингвистике, прежде всего, вычислительной семантики и приложениям, в которых применяются семантические языковые модели.

Материалы монографии распределены по принципу от «общего к частностям и деталям». Ее содержание начинается с аналитического обзора прагматически ориентированных лингвистических

моделей с точки зрения их использования как основы в системах и технологиях обработки естественного языка. Обзор, подготовленный Д. Ш. Сулеймановым, проведён в контексте опыта автора и его школы, полученного в разработках ряда прикладных лингвистически ориентированных систем.

Обзор начинается с анализа публикаций в области обработки естественного языка (более 100 наименований), нацеленного на обоснование прагматически-ориентированного подхода к разработке лингвистических моделей. Проводится сопоставление с традиционными решениями, используемыми в информационных технологиях обработки естественно-языковых текстов. Раскрываются идеи и модельные решения, близкие к прагматически-ориентированному подходу.

Важное внимание уделено анализу средств обработки естественно языковых текстов в диалоговых системах, с учетом особенности диалоговых моделей в аспекте указанного подхода.

Рассматриваются вопросы формализации семантической обработки текстов, специфицируется объектно-предикатная система как составляющая концептуально-функциональной модели, проводится анализ средств формального описания значений, а также анализ систем обработки текстов на основе концептуально-формальной модели.

Логико-лингвистическая прагматика нашла своё отражение в работах О. А. Невзоровой и П. И. Соснина, в которых раскрываются вопросы онтологического моделирования. В первой из этих публикаций рассматриваются базовые модели онтолого-лингвистических систем, ориентированых на решение сложных задач обработки естественного языка, требующих семантических знаний. В основе проектирования онтолингвистических систем лежат процессы скоординированного взаимодействия различных уровней, прежде всего онтологического (обеспечивающего системные модели знаний о мире) и различных языковых уровней. В работе предлагается концептуальная архитектура онтолого-лингвистических систем, а также решения по организации системы онтологических моделей в системе «ontointegrator»

В работе П. И. Соснина акцент смещён на персонификацию онтологичских моделей, обслуживающих процессы проектирования автоматизированных систем. Приводятся результаты исследований вопросов онтологизации персонального профессиональ-

ного опыта. Специфику решений по созданию и использование персональной онтологии определяет ориентация на представление опыта в виде системы моделей прецедентов, освоенных и созданных субъектом деятельности в реализации совокупности проектов. Модели прецедентов построены по образцу интеллектуально обработанных «деятельностных рефлексов» с использованием вопросно-ответных рассуждений.

Моделирование терминологических систем обсуждается в работе Л. Н. Беляевой, в которой рассматриваются лексикографические ресурсы переводчика с позиций их состава, структуры и ведения, предполагающих проведение предварительной терминологической работы для отбора и описания терминологии на разных языках, осуществление гармонизации этих описаний и согласование терминологических систем разных языков. Особо отмечается, что лексикографические ресурсы переводчика определяют оперативность, точность и корректность результатов его работы. Рассматривается исторический аспект разработки таких ресурсов, современные сетевые базы данных и методы их поддержки и ведения. Отдельно анализируются возможности использования лексикографических ресурсов в образовательной среде вуза.

Особую роль в вычислительной семантике выполняет кодирование и декодирование оценок различных ситуаций, вложенных в естественно языковые тексты. С таким классом задач в монографии связаны работы Н. В. Лукашевич и С. Г. Татевосова.

В первой из этих публикаций предлагается решение задачи анализа тональности (мнения, оценки), вложенной автором в текст, по отношению к заданному объекту, а также его характеристикам (аспектам). Показано, что для выявления тональности по отношению к аспектам сущности необходимо решать также задачи извлечения аспектов для сущности, категоризацию или кластеризацию аспектов по аспектным категориям, определение тональности текста по отношению к заданному аспекту или аспектной категории. Также в работе описываются открытые тестирования объектно-ориентированных систем анализа тональности.

В работе С. Г. Татевосова раскрыты подходы к предикатному моделированию свершений в привязке к четырем тюркским языкам (татарскому, чувашскому, алтайскому и карачаево-балкарскому). Рассматривается три подкласса свершений, различающихся тем, допускают ли они прочтение вида ‘неудавшаяся по-

пытка', 'частичный успех' или ни то ни другое. Автор исходит из положения, что существенной частью семантической структуры не кульминирующих предикатов является модальный оператор, выделяющий из событий их собственные неконечные стадии. В результате моделирования показано, что для успешного объяснения свойств не кульминирующих свершений необходим их декомпозициональный анализ, когда компоненты деятельности и изменения состояния становятся самостоятельными элементами семантического представления.

Таким образом, анализ проблем, затронутых в монографии, показывает широчайший спектр крайне интересных задач и приложений для новых исследователей и авторы надеются, что настоящая монография внесет свой посильный вклад в вычислительную лингвистику, в которой решения многих задач, связанных с формализацией семантики, еще только формируются.

П. И. Соснин, О. А. Невзорова

Глава 1. ПРАГМАТИЧЕСКИ-ОРИЕНТИРОВАННЫЕ ЛИНГВИСТИЧЕСКИЕ МОДЕЛИ КАК ОСНОВА СИСТЕМ И ТЕХНОЛОГИЙ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА. АНАЛИТИЧЕСКИЙ ОБЗОР

Д. Ш. Сулейманов

1.1. ВВЕДЕНИЕ

Актуальность проблемы. За более чем полувековую историю исследований и разработок в области автоматизации обработки естественно-языковой информации, задачи построения эффективных систем и технологий понимания текстов, естественно-языкового (ЕЯ) диалога с компьютером и систем с ЕЯ-интерфейсом продолжают оставаться по-прежнему актуальными и даже выходят в ряд наиболее острых и востребованных задач в связи с экспоненциальным ростом объема обрабатываемой информации, накапливаемой в различных базах знаний, особенно в связи с необходимостью поиска релевантной информации в глобальной сети Интернет.

В настоящее время можно констатировать, что, несмотря на определенные положительные результаты, сложные задачи обработки ЕЯ, требующие привлечения семантических знаний (семантический поиск, машинный перевод, системы обработки речи, ЕЯ человеко-машинный интерфейс, ЕЯ-диалог с компьютером) практически не решены на желаемом уровне.

Главной причиной такого положения, очевидно, является принципиальная сложность и междисциплинарность этих задач и то обстоятельство, что в основе соответствующих технологий, решающих эти задачи, лежит сложная семиотическая модель, которая концептуально объединяет элементы и структуры из различных предметных областей, таких как лингвистика (включая компьютерную и когнитивную лингвистику), информатика, психология, прикладная семиотика и математика.

В настоящее время среди ученых, занимающихся проблематикой автоматизации обработки естественно-языковой информации, есть понимание, что решение данной важной междисциплинарной задачи лежит в плоскости создания семантических технологий и интеллектуального программного инструментария, а также соответствующих семиотических моделей понимания естественного языка. Имеющиеся технологии можно охарактеризовать как линейные, позволяющие решать поставленные задачи скорее количественно, а не качественно, позволяющие объединять модели из различных проблемных областей (лингвистические модели, контекстные модели, модели "мира", целевые модели) последовательно-параллельно. Нужны нелинейные решения, нужны комплексные семиотические модели и технологии на их основе, иерархические и сетевые модели, в одной модели объединяющие различные объекты и структуры из различных проблемных областей.

Таким образом, ставится задача построения средств автоматизации обработки информации на основе принципиально новых представлений знаний, на основе концептуально иных семиотических знаков как единицы представления информации, на основе новых подходов и технологий.

В этой связи актуальной является разработка семиотических моделей, соответственно, систем и технологий обработки естественно-языковых текстов, достигающих эффективности за счет их ориентированности на определенный класс лингвистических задач.

Семиотические модели – это сложные представления, помогающие строить системы, которые обрабатывают: анализируют, генерируют, интерпретируют и трансформируют, естественно-языковые тексты, привлекая методологический, технологический, концептуальный и программный инструментарий из различных научных дисциплин, и наряду с лингвистическими моделями, также включающие экстралингвистические лингвистические модели, такие как модели «мира», модели контекста, модели принятия решений.

Очевидно, построение систем и информационных технологий на основе универсальной лингвистической модели является малоперспективным, с одной стороны, ввиду отсутствия универсальной формальной базы, единой теории и формального инстру-

ментария описания языков, и даже приемлемой полной формальной модели какого-либо языка, с другой стороны, из-за того, что реализация даже некой модели, приближенной к универсальной, с применением современных технологий, будет неэффективной и малопригодной по временным и емкостным характеристикам.

Известные системы обработки ЕЯ-текстов, как правило, созданы на основе лингвистической модели, включающей полный или ограниченный анализ морфологии, синтаксиса и семантики без учета многих прагматических факторов, в частности, без учета специфики классов лингвистических задач. Эффективность таких систем достигается за счет существенных изначальных ограничений либо на язык, либо на проблемную область, либо на обе эти составляющие.

Нами разработан подход посторения лингвистических процессоров на основе прагматически-ориентированных моделей, или моделей от цели, позволяющих перейти от технологии построения систем путем введения ограничений на универсальную языковую модель к технологии разработки прагматически-ориентированных, целевых моделей, включающих изначально минимальный набор средств, разработанных исходя из принципа достаточности для решения определенной задачи.

Разработанные нами 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, что достаточно полно отражено в работах автора данной статьи.

В данной работе представлен анализ существующих лингвистических моделей, систем и информационных технологий обработки ЕЯ-текстов; исследованы современные подходы и методы построения лингвистических моделей, систем и информационных

технологий обработки естественно-языковых текстов, а также определены принципы и критерии их построения; исследованы грамматические формализмы как основа лингвопроцессоров с точки зрения их результативности и эффективности; сформированы теоретические положения прагматически-ориентированного подхода и понятийный аппарат, необходимые для описания прагматически-ориентированных лингвистических моделей; исследованы технологии создания и применения ЕЯ-систем.

В работе, в большей степени, мы ограничимся рассмотрением описаний систем и аспектов, недостаточно раскрытых в имеющейся обзорной литературе.

Разработана методология и предложены методы и базовые принципы построения систем 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.

1.2. ПРАГМАТИЧЕСКИ-ОРИЕНТИРОВАННЫЙ ПОДХОД К РАЗРАБОТКЕ ЛИНГВИСТИЧЕСКИХ МОДЕЛЕЙ

Имеется немало обзорных аналитических работ, посвященных предыстории, тенденциям развития ЕЯ-систем, средствам описания компьютерного представления и обработки лингвистической информации. К таким работам, достаточно полно характеризующим предысторию развития, современное состояние и тенденции в области обработки естественно-языковых текстов, мы относим фундаментальные исследования Мальковского М. Г., Ronald A. Cole, А. С. Нариньяни, С. А. Шарова, Э. В. Попова, Дж. Слокум, Б. Ю. Городецкого, С. О. Шереметьевой.

При анализе тенденций развития работ в этой области исследователи, как правило, придерживаются схожей стратегиче-

ской линии и выделяют следующие три подхода, отмеченных 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. Первыми формальными средствами, развитыми для описания языков, наиболее исследованными теоретически, являются грамматики Хомского. Поэтому вполне закономерно, что обзор развития лингвистической теории, включает, как правило, трансформационные грамматики Н. Хомского, описывающие два уровня синтаксической структуры (глубинной и поверхностной), связанные посредством трансформаций, ставшие фундаментальной идеей для многих дальнейших исследований и реализаций; модель «Смысл <-> Текст», в которой язык рассматривается как система кодов, соответствующей системе смыслов; а также наиболее популярную в настоящее время в компьютерной лингвистике современную модель Хомского GB, реализующую принцип ограничений на сформированность лингвистической структуры для описания грамматики. Идея генеративной грамматики Н. Хомского, «которая в своем современном виде включает в себя многие достижения функциональных теорий языка последних лет», и процесс их исторического развития подробно изложены Дж. Бейлин в сборнике обзоров «Фундаментальные направления современной американской лингвистики».

Однако грамматики Хомского, порождающие всевозможные правильные синтаксические конструкции, по мощности избыточны для анализа синтаксиса, и в то же время, недостаточны и скорее даже принципаильно не подходят для установления семантических аспектов фразы на естественном языке.

В связи с этим появились грамматики для анализа текстов, являющиеся модификацией или расширением КС-грамматик, позволяющие учитывать семантические признаки. К таким грамматикам можно отнести и грамматическую теорию HPSG (Head-Driven Phrase-Structure Grammar), в которой знак представляется как набор атрибутов различных уровней, позволяющий интегрировать всю лингвистически-релевантную информацию от фонологической до прагматической с одновременным ее использова-

нием; грамматику SFG (Systemic Functional Grammar), основным понятием которой является система как набор возможностей для некоторого грамматического признака; и грамматика LFG (Lexical Functional Grammar).

Экспериментальный подход к построению лингвистических систем, как правило, используется специалистами в области искусственного интеллекта в связи с тем, что стремление построить реально функционирующую эффективную систему понимания естественного языка и неудачи на этом пути привели к осознанию того, что глубокое проникновение в содержание ЕЯ-текста и рассматриваемого контекста возможно только в рамках четко определенной узкой проблемной области и при достаточно жестких ограничениях на язык.

Обзор экспериментальных систем обработки ЕЯ-текстов, как правило, содержит описания семантических падежей, семантических ролей, определяющих роли участников ситуации, управляемых предикатом; модель «семантик предпочтения», рассматривающей смысл предложения не просто как список значений слов с соответствующей синтаксической структурой, а выделяющей структурированную форму сообщения, выражающую смысл предложения; модель концептуальной зависимости, предназначенной по замыслу авторов для описания модели мышления человека, характеризующейся преобладанием роли семантики и, практически, отсутствием привычных представлений синтаксиса и морфологии. Такая интерпретация слов проста и удобна, поскольку позволяет получить о функциях слов, участвующих в каждом предложении, полную картину такого вида: *кто, что делает, по отношению к кому, с кем, когда, где* и др. в зависимости от того, какие роли существенны для выбранной языковой подсистемы. Это обеспечивает достаточно простую процедуру семантического анализа на логико-ситуационном глубинном уровне.

Естественно, такая обработка текста, основанная на упрощенной модели языка, позволяет игнорировать сложности синтаксиса и создает хорошие предпосылки построения достаточно реактивных анализаторов текста. Однако, практически все существующие модели, как уже было сказано, носят экспериментальный характер, и, несмотря на многие положительные качества, неприемлемы для их прямого использования в прагматически-ориентированных лингвистических моделях, в таких как вопросно-ответные

диалоговые модели, имеющие ряд выгодных особенностей по сравнению с универсальными моделями обработки ЕЯ-текстов (детально исследуется в работах Сулейманова Д. Ш.).

Прагматический подход в литературе, как правило, ассоциируются с созданием реальных система, созданными для решения конкретного круга задач в узкой проблемной области (ПО), и достигающих высокой эффективности за счет ограниченности ПО, однозначности контекста и примитивности языкового интерфейса. При этом прагматика, т.е. нацеленность на результат, проявляется в процессе эксплуатации пользователем готовой системы.

Отдельного рассмотрения заслуживают работы Нариньяни А. С., отчасти отраженные в материалах конференций ДИАЛОГ и КИИ, отличающиеся оригинальностью и новизной, касающиеся ситуации в области автоматической обработки текста, интеллектуализации информационных технологий, аппарата знаний и новых поколений приложений.

Будучи одним из пионеров исследований в области компьютерной лингвистики и активным разработчиком современных систем понимания ЕЯ-текстов, Нариньяни А. С. критически оценивает состояние в области создания лингвистических моделей, подчеркивая смещение акцентов в сторону бесперспективного алгоритмического подхода, и утверждает, что «следующее поколение технологии автоматической обработки текста должно ориентироваться на принципы, в корне отличные от тех, на которых до последнего времени базировалось подавляющее большинство проектов в этой области». Оценивая результаты более трех десятилетий исследований и разработок в области автоматического понимания ЕЯ текста А. С. Нариньяни в одной из работ отмечает, что «результатов до обидного мало... в рамках общепринятой до сегодня методологии никакая прикладная задача, включающая понимание ЕЯ текста, не может быть решена по крайней мере в ближайшие десять (а скорее всего, и тридцать) лет».

Современное состояние дел с системами и технологиями в области обработки ЕЯ-текстов, особенно в плане развития семантических технологий, показывает, насколько пророческими были слова А. С. Нариньяни, одного из основателей и исследователей научно-прикладного направления компьютерная и математическая лингвистика в нашей стране. Усилиями А. С. Нариньяни и А. Е. Кибрика в СССР, а потом в России была организована

международная конференция по компьютерной лингвистике «Диалог», традиции проведения которой в настоящее время на высоком научном и практическом уровне продолжает фирма ABBYY.

Одной из наиболее интересных работ, представляющей некий целостный механизм для инженерии языка (Language Engeneering – LE), является проект GATE (a General Architecture for Text Engineering), отражающий продвинутое представление об общей архитектуре систем обработки текстов. Проект GATE является архитектурой, обеспечивающей общую инфраструктуру для разработки LE-систем, и содержит три основных модуля:

– база данных для хранения текстовой информации и оболочка базы данных, основанная на объектно-ориентированной модели (the GATE Document Manager – GDM);

– графический интерфейс для запуска средств обработки данных, просмотра и оценки результатов (the GATE Graphical Interface – GGI);

– совокупность объектов для ресурсов алгоритмов и данных, которые взаимодействуют с базой данных и интерфейсом и образуют совокупность повторно используемых объектов для задач LE (a Collection of Reusable Objects for Language Engineering – CREOLE).

Однако, оценивая положительно создание единого комплексного механизма обработки текстов, удобного для разработчика и пользователя при проектировании систем и их эксплуатации, необходимо все же отметить, что такое агрегирование и технологизация, практически, не привносят ничего принципиально нового в идеологию и методологию LE. Более того, появляются трудности другого плана, связанные с несовместимостью представления текстовой информации с механизмами хранения, извлечения и межмодульного взаимодействия и несовместимостью типов информации в различных модулях.

В настоящее время имеется достаточное число подробных обзорных работ в области систем обработки ЕЯ-текстов, а также материалы семинаров и конференций, посвященных проблемам обработки ЕЯ-текстов, в которых исследуются современные подходы и методы их построения. Учитывая данное обстоятельство, в настоящей работе нами выполнен аналитический обзор работ, в которых высказываются идеи, или предложены раз-

работки, близкие к сформулированному нами прагматически-ориентированному подходу в построении лингвистических моделей и реализации систем обработки ЕЯ-текстов.

Несмотря на обилие научной и технической литературы, посвященной описанию лингвистических процессоров, от узкоспециализированных до универсальных, ни одна из приведенных задач в настоящее время, практически 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. И причина здесь, как нам представляется, заключается в том, что, во-первых, как правило, модели строятся не «от задачи», а «от языка», и, во-вторых, реальность всегда оказывается сложнее, нежели предполагается вначале, сколь бы подробно ни описывалась модель языка.

Построение лингвистических моделей исходило изначально из сомнительного утверждения, что «для того, чтобы допускать возможность реальной компьютерной реализации, лингвистическая теория должна обладать высокой степенью формализации и полноты», т.е. делалась попытка построить идеальную инструментальную систему обработки ЕЯ, основанную на идеальной лингвистической теории. Попытка найти решение через создание новых формализмов, ориентированных на описание лингвистических феноменов (морфологии, синтаксиса, семантики и др.), после каждой неудачи с предыдущими моделями, привела к большому многообразию лингвистических моделей, практически, различающихся только набором средств описания лингвистических феноменов, но не результативностью.

Решение проблемы построения *естественно-языковых лингвистических моделей* и создания на их основе *эффективных лингвопроцессоров*, как нам представляется, 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. К числу участников

обработки текстов относятся все привлекаемые ресурсы и субъекты, включая обрабатываемый ЕЯ-текст, пользователя, нелингвистические (экстралингвистические) блоки системы, проблемную область, контекст и т.п.

Под обработкой ЕЯ-текстов понимается процесс взаимодействия «Система-Текст-Пользователь», включающий различные способы воздействия на текст, такие как анализ, генерация, интерпретация, трансформация и др. Такое определение семиотических моделей, основанное на их функциональном аспекте, является весьма продуктивным с методологической точки зрения, и позволяет провести соответствующую классификацию моделей по их прагматическим признакам (по цели разработки и сфере применения), а также разработать достаточно эффективные технологии и системы обработки ЕЯ-текстов. Нами проведена следующая ***классификация лингвистических моделей***:

1) семиотические (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 соответствующим формальным правилам определенного языкового уровня (согласно собственно грамматике ЕЯ);

4) концептуально-функциональные модели, являющиеся наиболее полными описаниями элементов (значимых единиц) определенного ЕЯ-уровня или уровней, обеспечивающими разработчиков структурно-функциональной, а также справочной информацией, необходимой при построении систем и информационных технологий обработки естественно-языковых текстов (например, полное описание аффиксальных морфем в их проявлении на всех языковых уровнях – морфологии, синтаксиса, семантики, фонетики).

Первые три класса моделей, главным образом, отличаются сложностью моделей, т.е. количеством языковых уровней и связей между ними и средствами экстралингвистического описания. Самыми сложными, очевидно, являются модели, относящиеся к первому классу моделей, включающие наибольшее число различных взаимосвязанных языковых уровней (в общем случае – все уровни языка), а также экстралингвистические описания, характеризующие особенности речевого акта, такие как модальность, неполнота, контекст, модель «мира», пропозиция, иллокутивные и коммуникативные особенности и др. Эти модели являются наименее изученными и слабо описанными в настоящее время.

Для моделей, относящихся к первому классу, определяющим является привлечение глубинных, когнитивных и семантических представлений, достаточно адекватно описывающих проблемную среду и контекстные проявления языка, не поддающиеся полной формализации.

Наименее сложными, поддающимися максимальной формализации, но емкими и требующими максимальной полноты описания языкового уровня, являются модели, относящиеся к третьему классу моделей.

Модели второго класса являются наиболее разработанными и представленными в экспериментальных и практических реализациях, занимая по сложности и определенности описаний серединную позицию между моделями первого и третьего классов. К обработке естественно-языковых текстов семиотические модели второго класса, по необходимости, могут привлекаться как модели с элементами глубинных представлений, так и формальных описаний естественного языка. Главной отличительной особенностью моделей второго класса, т.е. диалоговых моделей, является то, что они обязательно содержат специальные блоки формирования образа ответа на запрос или реакцию на ответ.

Лингвистические модели четвертого класса, практически являются «инвентарными» моделями, 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.

Из сказанного следует, что приведенная классификация не является вложенной, т.е. ни один из классов не является частью

другого. Каждый класс моделей определяется наиболее характерной и максимально полно описываемой в ней составляющей, что, естественно, не исключает подключения моделей из других классов при разработке лингвистических процессоров.

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, 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 методологичсеких принципов «ожидаемости» и «детерминированости 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 ***«детерминированости контекста»*** естественным образом «диктует» параметры входного текста – его содержание, форму, лексикон. Очевидно, контекстное управление является естественным для диалоговых моделей, когда один из участников взаимодействия, а именно, обрабатывающий текст, владеет инициативой. В случаях лингвистических моделей первого и третьего классов, как правило, тексты являются заданными и не зависящими от участника, обрабатывающего его, т.е. обрабатывающая сторона уже не является активной. Тем не менее, в силу того, что диалоговые модули в этих классах моделей, зачастую, используются в качестве служебных, принцип «детерминированости контекста», хотя и опосредованно, здесь также имеет место.

Благодаря принципу «ожидаемости», текст также становится активной стороной процесса обработки в рамках своей «компетенции», т.е. «заставляет» процессор «мобилизовать» целенаправленно те ресурсы, которые нужны для обработки текстов именно такого, определенного, типа.

Аналогично, «ожидаемость» определенной ситуации взаимодействия (текст для перевода, для извлечения знаний, диалоговый текст, текст для морфологической обработки и т.п.) также способ-

ствует минимизации и «опережающей» перегруппировке средств, привлекаемых для обработки текста.

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 стороны, 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 процедурного, 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 перефразироввания, для извлечения знаний, вопрос-ответ, запрос-ответ, для морфологического анализа, морфологической коррекции и др.), следовательно, определяется минимальная часть средств, выбранных пунктом (а).

Очевидно, учет специфики выделенных классов моделей, а также возможная редукция и дифференциация средств внутри этих классов с учетом более тонких различий, позволяют повысить эффективность лингвопроцессоров, построенных на их базе. Основная идея прагматически-ориентированного подхода к построению лингвистических моделей, хотя и напоминает в некоторых аспектах другие подходы к анализу ЕЯ, представляет собой новую и оригинальную методологию, предложенную и воплощенную автором статьи в прикладных системах.

В зависимости от цели моделирования, может быть реализовано множество различных прагматически-ориентированных лингвистических моделей в рамках выделенных четырех классов. В наших работах исследованы и построены лингвистические модели, соответствующие классам моделей 2-4. В связи с этим мы сочли целесообразным провести целевой сравнительный анализ работ, касающихся диалоговых, концептуально-формальных и концептуально-функциональных моделей, отталкиваясь именно от разработанных и описываемых нами моделей.

Подробный обзор систем, относящихся к первому классу, т.е. классу когнитивных моделей, приводится в работах Поспелова Д. А. и Полякова В. Н.

1.3. АНАЛИЗ БЛИЗКИХ ИДЕЙ И ПОДХОДОВ

В работах Нариньяни А.С. раскрываются четыре следующих взаимосвязанных и взаимодополняющих принципа:

1) Семантически-ориентированный подход к анализу ЕЯ текста.

2) Эффективное использование знаний (как проблемных, так и общих) во время и после лингвистических этапов процесса анализа-понимания.

3) Организованное сообщество активных *constraint-based* агентов, а не система продукционных правил в качестве аппарата спецификации лингвистического обеспечения.

4) Снизу-вверх и распределенная, а не сверху-вниз централизованная организация процесса обработки, базирующегося на управлении по данным и\или событиям, а не на традиционном императивном типе управления.

Такой подход во многом напоминает предложенный нами прагматически-ориентированный подход, однако, каждый из них имеет свою специфику, и скорее, эти подходы, совпадая в основе, в ряде моментов дополняют друг друга. Так, семантически-ориентированный подход к анализу ЕЯ текста определяется А. С. Нариньяни следующим образом: «*пытайся восстановить смысл текста, используя всю доступную семантическую и прагматическую информацию; обращайся к синтаксическим компонентам только тогда, когда это необходимо для разрешения неоднозначности; это обращение должно соответство-*

вать требованию минимальной достаточности – используй минимум информации, нужной для решения данной локальной задачи». Технология прагматически-ориентированного подхода также включает в качестве составляющих практически те же этапы, однако, согласно определению, она используется, начиная с этапа методологии построения лингвистической модели, т.е. до того, как начинают фигурировать в качестве звена технологии сам текст и семантическая информация по нему.

Второй принцип, т.е. эффективное использование знаний, соответственно, также расширяется долингвистическим этапом, а третий и четвертый принципы, вполне приемлемые в моделях «понимания» типа семиотических и диалоговых (типа запросно-ответных или гибких), выглядят слишком категоричными для концептуально-формальных, концептуально-функциональных и диалоговых вопросно-ответных (жестких) моделей, выделяемых по прагматически-ориентированной технологии.

В целом ряде проанализированных нами работ, авторы также явно или неявно высказывают, или даже используют при разработке конкретных систем, идеи, близкие по содержанию к идее прагматически-ориентированного подхода. Так, в работе Андрусенко Т. Б. приводится следующее определение: *«...прагматически ориентированным можно считать направление прикладных исследований и разработок, целью которых является определение теоретических и практических основ создания систем диалога с ЭВМ с учетом особенностей решаемых задач и стратегии использования ЕЯ».* Как справедливо замечает Андрусенко, прагматика даже в таком более узком ее понимании по-прежнему представляет собой наименее разработанный аспект диалога, а для более общих случаев построения прагматически-ориентированных моделей и технологий вопрос в такой постановке, практически, в литературе не рассматривается. Исследователи Дж. Гвида и К. Тассо предлагают подход к созданию систем понимания естественных языков, называемый целенаправленным анализом, подтверждая, тем самым, объективность и перспективность предлагаемой нами технологии построения прагматически-ориентированных лингвистических моделей. Под ЕЯ-интерфейсом, как правило, декларируется запрос к системе. В нашем же случае, естественно-языковой текст – это ответ на вопрос системы, причем, естественность заключается в том, что

не накладывается никаких специальных ограничений на форму и полноту языка ответа.

Ранее пути решения общей проблемы понимания ЕЯ базировались, в основном, на чисто лингвистическом анализе входного текста. Однако, очевидно, что оптимальную модель понимания ЕЯ можно построить, если только эксплицитно учитывать ее цели и сферу применения. Если во главу угла будет поставлена цель взаимодействия с ЭВМ на ЕЯ, то это даст возможность отвлечься от многих лишних деталей и тонкостей входного текста и выбирать из них только релевантную информацию. Такой подход позволит увеличить эффективность алгоритмов анализа текста и в то же время обеспечит свободное взаимодействие с ЭВМ на ЕЯ. Ряд исследователей замечает, что такой принцип целенаправленного (goaloriented) анализа может с успехом использоваться в общей лингвистической теории, в которой семантика ЕЯ в контексте взаимодействия человека с ЭВМ определяется на основе таких понятий, как поведение, цели и планы слушающего и говорящего.

По мере накопления опыта разработки и эксплуатации систем обработки текстов прагматическая концепция постепенно складывается как требование исследования и реализации лингвистических моделей различного назначения. Одним из аспектов такого исследования может служить попытка сформулировать задачу следующим образом: *определение влияния так называемых неязыковых факторов общения на выбор языковых средств обработки сообщений. Этими средствами выступают как целые языковые структуры (уровень композиционной схемы диалога), так и конкретные языковые элементы.*

На необходимость комплексного, технологичного подхода к исследованию проблемы «понимания» текста, большей интеграции между семантикой уровня предложения или высказывания и теориями структур текста или диалога указывается и исследователем Pulman Stephan G. Semantics. Как подчеркивается в работах Ефимова Е. И., опыт разработки систем понимания и синтеза связных текстов показывает, что многие трудности, как правило, вызываются отсутствием указания на практическое приложение данной проблемы. Отсюда возникают неопределенности, связанные с описанием внешней среды диалога, его темы и целей, описанием внутренних миров партнеров и механизмов генерации, восприятия и когнитивной обработки речевых актов.

Некоторые важные идеи раскрываемого нами прагматически-ориентированного подхода к разработке систем обработки ЕЯ-текстов высказаны и использованы при создании конкретных систем в работах Мальковского М. Г. Здесь подчеркивается, что «переход к созданию прикладных систем общения с ЭВМ на ЕЯ ставит перед разработчиками целый комплекс вопросов, не возникавших на этапе первых экспериментов. Серьезная практическая задача обеспечения взаимодействия с машиной на ЕЯ требует серьезного и практического подхода». Акцентируя внимание на прагматике и приводя основательные доводы в пользу ЕЯ-диалога, Мальковский М. Г. также подчеркивает разумность выбора пути на ограничения ЕЯ. Если диалоговая система узко ориентирована (узка ее область применения, функции), разумнее, вероятно, использовать не так называемый «ограниченный ЕЯ», а специализированный формальный язык общения.

Анализ существующих лингвистических моделей и тенденций в этой области показывает актуальность и естественность разработки прагматически-ориентированной технологии построения лингвистических моделей систем и информационных технологий на их основе, что, как мы полагаем, и демонстрирует также и опыт построения реальных моделей обработки ЕЯ-текстов, описанных в работах автора данного раздела монографии.

1.4. АНАЛИЗ СРЕДСТВ ОБРАБОТКИ ЕЯ-ТЕКСТОВ В ДИАЛОГОВЫХ СИСТЕМАХ (В АСПЕКТЕ ПРАГМАТИЧЕСКИ-ОРИЕНТИРОВАННОГО ПОДХОДА)

В последние годы активизировались теоретические и прикладные работы по развитию комплексной и многоаспектной лингвистической стороны проблемы человеко-машинного диалога, а именно, исследование диалога как способа общения и вида текста.

Прагматически определяющим фактором человеко-машинного диалога является то, что она формируется на основе связи темы диалога, ситуации и контекстов взаимодействия в зависимости от коммуникативных задач, стоящих перед участниками общения, а также языковых возможностей, которыми обладают участники диалога. Как утверждается в работах Андрусенко Т. Б., формализация процесса общения в целях построения систем

человеко-машинного диалога (ЧМ-диалога) диктует определение таких языковых структур, которые могут быть порождены и интерпретированы алгоритмически. Факторы, влияющие на выбор языковых средств и формирование текста ЧМ-диалога: *фактор подготовленности диалога, тематически ограниченная предметная область, факт структуризации предметной области (ПО) в сознании человека и в памяти системы, стереотипный характер ситуаций ЧМ-диалога, предсказуемость и ожидания в стереотипных ситуациях.* К главным критериям выделения предпочтительных языковых средств Андрусенко Т.Б. относит типы проблемных ситуаций задачи, решаемой в диалоге, и способность языковых элементов выступать в роли того или иного компонента проблемной ситуации.

Установление объективности категории предпочтительных языковых средств выражается в концепции ограниченного естественного языка как варианта языка диалога, который является ограничением не по своей природе, а по степени использования тех или иных средств передачи сообщений. Далее рассматриваются признаки и отношения, характерные для текста человеко-машинного диалога. Наиболее структурированный уровень отношений составляют отношения между элементами структуры диалога и языковыми средствами. Являясь отражением отношений между фрагментами действительности (зафиксированными в некоторой модели представления знаний о ПО) и описывающими их языковыми средствами (составляющими материальную основу текста), что выражается категорией смысла, они образуют как бы серединный уровень следующей иерархии:

действительность	**смысл (отношение)**	**язык**
Модель действительности (структуры знаний о ПО)	Смысл – соответствие	языковые средства (текст)
Структура диалога (комплекс диалогических элементов и отношений)	Смысл – прогнозирование	языковые средства (фрагменты теста)
Ситуация и элементы ситуации	Смысл-функция	языковые средства (конкретные языковые элементы)

В первом аспекте смысл текста в ЧМ-диалоге рассматривается в связи с категориями цели и функций участников при решении некоторой задачи. В этом случае понимание каждым участником диалога языкового выражения связывается с выполнением своей функции на данном этапе диалога, вытекающей из содержания анализируемого выражения. Например, может быть установлена следующая ситуация: ***тип диалога*** *– вопросно-ответный;* ***режим диалога*** *– второй, т.е. активна система (ведущая), пассивен ученик (ведомый).*

Второй аспект смысла ассоциируется с категорией прогнозируемости стереотипных ситуаций и их последовательностей, а также закрепленных за ними устойчивых языковых форм, т.е. понимание в этом случае связывается с прогнозированием. В нашем случае, это – определение типа вопросов и ответов и концептуальной грамматики как совокупности ИКГ.

В третьем аспекте смысл выступает как отношение между моделью действительности и текстом, где текст фиксирует некоторое положение вещей в мире. Здесь понимание связывается с соответствием содержания языкового выражения данному фрагменту действительности. Категория соответствия включает как результат действия на основе понимаемого (воспринимаемого) смысла, так и его оценку, которая завершает текущий цикл понимания. Такая ситуация соответствует анализу на уровне модели значений заданного вопроса и соответствующей индивидуальной концептуальной грамматики.

Отдельным типом отношений между элементами и единицами диалога можно считать тематическую связность. Тема диалога всегда известна и структурирована в модели предметной области. Для вопросно-ответной ситуации, средством, обеспечивающим тематическую связность, выступают типы вопросов и классы значений вопросов, т.е. «ожидаемых» ответов обучаемого.

Формально-грамматическая связность базируется на конкретных языковых элементах и отношениях, служит основой формирования текста. В диалоговой модели формально-грамматическая связность реализуется через индивидуальные концептуальные грамматики каждого класса значений вопросов, связывающие глубинные смысловые конструкции анализируемого текста с ожидаемой моделью значений.

В лингвистическом процессоре ЛП, разработанном на основе прагматически-ориентированной диалоговой модели, минимальной коммуникативной единицей выступает текст, создаваемый в вопросно-ответной ситуации, т.е. ответ обучаемого на вопрос системы, являющийся одним из ожидаемых значений вопроса (в отличие запросно-ответной ситуации, в которой текст является либо запросным текстом пользователя к системе, либо реакцией (ответом) системы на некоторый запрос пользователя).

Запросно-ответные системы

Диалог человека с машиной означает интерактивный обмен посланиями между пользователем и диалоговой системой в соответствии с условленными языком (языками) диалога и формой (формами) диалога для достижения определенной задачи. 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, и наконец, 3) двухсторонне активный диалог, когда пользователь и система меняются ролями в ходе общения.*

Наиболее изученным, развитым и представленным в литературе является режим (2), когда вопросы задает пользователь, а система отвечает. Очевидно, что при этом успешный поиск информации в базе данных и генерирование соответствующего ответа, главным образом, зависят от того, насколько корректно система сможет интерпретировать вопрос пользователя. Большинство лингвистических процессоров для общения с базой данных (БД) на ЕЯ, активно разрабатываемых рядом отечественных и зарубежных исследовательских групп и фирм, не опираются на принципиальную лингвистическую модель и функционируют в предположении, что человек инициирует диалог, т.е. цели пользователя, а не системы, определяют диалог. К таким работам относятся экспериментальная система TIBAQ (Text-and-Inference Based Answering of Questions)), а также исследования Н. Белнапа и Т. Стила, У. Ленерта, системы SAM (Р. Шенк и др.), ПОЭТ (Э. В. Попов и др.), многие экспертные и информационно-справочные системы, например, MYCIN (Шортлифф), ИВОС

(Л. Т. Кузин, А. Б. Преображенский, В. Ф. Хорошевский и др.), Лингвистический процессор для сложных информационных систем (Ю. Д. Апресян и др.), система InterBASE (А. С. Нариньяни и др.).

Как показывает анализ систем, ориентированных на запросно-ответный диалог, каждая из них, имея свои особенности, реализует следующую общую схему: воспринимает сообщение пользователя как запрос и формирует соответствующий ответ на основе знаний.

В работах Мальковского М. Г. рассматриваются аспекты естественно-языкового запросно-ответного интерфейса в обучающей ситуации. Обоснованно утверждается, что автоматизированное обучение является одной из таких сфер, где использование для общения с ЭВМ ЕЯ целесообразно и с точки зрения удобства для пользователя, и по существу.

Аннотированный библиографический указатель работ по теории вопросов и ответов (составители У. Эгли и Х. Шлейхерт), содержащий ссылки на более чем 200 источников с аннотациями, также демонстрирует интенсивность исследований и разработок именно для второго подхода, т.е., когда активным является пользователь. Классификация вопросов в большинстве случаев осуществляется: либо по лексическому принципу (например, *что-, почему-, как*-вопросы), либо по степени конкретности запроса и эксплицитности или имплицитности поисковой информации в базе данных.

Выше рассматривались системы, функционирующие в ситуации, когда активен пользователь. Диалоговое общение с системой в режимах 1 и 3 практически остается малоизученной областью, что вполне объяснимо. Режим 1 в чистом виде встречается на практике реже, чем второй режим взаимодействия человек-ЭВМ, и в большей степени, этот режим изучается как часть режима 3, когда осуществляется общение с ЭВМ с переменной инициативой участников общения.

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.

Рассмотрению вопросно-ответных текстов и построению соответствующих формализаций именно в такой ситуации посвящены работы Сулейманова Д. Ш.

Вопросно-ответные системы

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 (АОС). Вопросно-ответная ситуация в АОС имеет следующие особенности, учет которых позволяет строить прагматически-ориентированные лингвистические модели как основы эффективных анализаторов ответов обучаемого.

1) Особенность входного текста. В АОС текст на ЕЯ – это ожидаемый ответ на заданный вопрос, иначе говоря, это множество значений заданного вопроса. Вопрос АВТОРа накладывает определенные ограничения на форму ответа и его содержание. Ожидаемый объем ответа ограничивается требуемой степенью подробности по заданному вопросу. В силу определенности контекста и ожидаемых лексем, сводится к минимуму неоднозначность лексем (омонимия, полисемия, синонимия и др.).

2) Особенность «понимания» текста. Задача анализа ответов в АОС – это проверка правильности ответа обучаемого, т.е. соответствия его ожидаемому. При этом в АОС, зачастую, достаточно извлечение из текста только той его части, где содержится ответ на заданный вопрос. В качестве такой части может быть выделен также некоторый текст на ЕЯ. Выделенная часть текста подвергается лингвистической обработке (возможно, в специально определенных терминах).

3) Особенность семантической классификации текстов. Многообразие форм представления смысла текста на ЕЯ требует определения принципов элиминации этой многозначности на основе типовых смысловых конструкций. Одним из таких принципов является принцип семантической классификации вопросно-ответных текстов. Семантическая типизация вопросов позволяет разбить множество ответов обучаемого на семантические классы, в каждом из которых требуется раскрытие некоторого однотип-

ного смысла, определенного типом вопроса и независимого от формы задания вопроса.

Смысловая типизация вопросов и семантическая классификация ответных текстов дают возможность противопоставить каждому типу вопроса ограниченный набор допустимых, т.е. логически правильных, смысловых конструкций (формул). Можно рассматривать совокупность этих формул, соответствующих конкретному типу вопроса, как некоторую грамматику, кодирующую конструкции, передающие правильный смысл ответа.

4) Особенность формальной основы анализа. В существующих анализаторах ответов обучаемого в АОС грамматика естественного языка либо полностью игнорируется, либо используется жесткая «грамматика» в рамках строго формализованного обучающего курса.

Диалог обучаемого с АОС предполагает вопросно-ответную ситуацию, когда задающему вопрос (т.е. АОС) естественно ожидать в ответе обучаемого раскрытия определенного смысла, заданного вопросом, ограниченного объема лексем в ответе, с большой точностью соответствующих лексемам, ожидаемым по заданному вопросу. При этом важно не поверхностно-синтаксическое различие фраз, а то, какое предметное содержание имеет слово в ответе. Это содержание не зависит ни от части речи, которой выражено слово, ни от того, каким членом предложения оно является, а определяется той ролью, которую выполняет соответствующая лексема в ряду других в текущем контексте. Вследствие этого, при контроле ответа обучаемого в АОС для получения эффективных алгоритмов анализа ЕЯ-текста могут быть использованы упрощенные лингвистические модели, ориентированные на информированного (т.е. знакомого с контекстом) «слушающего».

5) Особенность выходной информации. В результате анализа ответов обучаемого необходимо получить набор параметров, характеризующих степень правильности ответа (диагностику), с целью управления учебным процессом. Как известно, управление процессом обучения зависит от многих параметров, в том числе от дидактических требований, опыта преподавателя, особенности каждого предмета, предыстории обучения и т.п.

Как показывает анализ возможностей контроля естественно-языковых текстов в реальных обучающих системах, описанных в ряде работ, практически, ни одна из исследованных разработок

не содержит модуль, предназначенный для анализа естественно-языковых ответов обучаемого по смыслу. Подробный сравнительный анализ средств автоматизации контроля ответов обучаемого приводится в совместной монографии Сулейманова Д. Ш. с Бухараевым Р. Г. Задача семантического анализа естественно-языковых текстов в проанализированных нами традиционных АОС не решена, более того, не существует конструктивных методов формального определения семантики. Анализатор ответов обучаемого, включенный в состав лингвистического обеспечения АОС ВУЗ-ОСКАР, развитый возможностями анализа ответов новых типов вопросов и семантических классов значений вопросов реализует частное решение задачи семантического анализа ответных текстов на естественном языке на основе прагматически-ориентированного подхода.

1.5. ФОРМАЛИЗМЫ В ОСНОВЕ СИСТЕМ СЕМАНТИЧЕСКОЙ ОБРАБОТКИ ЕЯ-ТЕКСТОВ

В современных исследованиях по компьютерной лингвистике имеет место определенная поляризация. С одной стороны, разработаны очень простые грамматические модели, т.е. различного рода грамматики конечного состояния, которые поддерживают высокую эффективность обработки. Некоторые подходы отказываются от грамматик и используют статистические методы для основных лингвистических шаблонов. С другой стороны, разработан целый ряд представлений мощных и лингвистически сложных формализмов, предназначенных для построения грамматик.

Наиболее распространенная группа грамматических формализмов, используемая в настоящее время в вычислительной лингвистике – формализмы на основе ограничений.

С. Пулман в своих работах, исследует «извечную» проблему семантики – выделение ее главного элемента – составляющего, сути. Семантика здесь понимается как буквальная интерпретация предложений в контексте, не принимая в расчет такие феномены как ирония, метафоры, или разговорные недомолвки. Утверждается, что знание значения предложения может быть приравнено знанию условий его корректности: т.е. знанию, которое порождает впечатление, что исследуемое предложение коррек-

тно в этом мире. Очевидно, это не то же самое, что знать, является ли предложение верным (корректным), что есть вопрос эмпирический, а то, что знание условий корректности есть предпосылка, чтобы была возможна их проверка. Признание ***значения*** как совокупности корректных условий, которое, практически, в той или иной форме является общим для всех современных теорий и имеет свое философское обоснование, также нуждается в обобщении каким-то образом и для императивов и вопросов. Семантическое описание языка есть некоторый конечно-установленный механизм, который позволяет для каждого предложения утверждать, какие условия для него являются корректными.

Язык для записи глубинной структуры и некоторые правила перевода глубинных структур в поверхностные предложили ряд исследователей, в их числе Ч. Филлмор, Р. Шенк, П. Уинстон и др. Между «ролями» (элементами глубинной структуры), и аргументами (элементами текста и элементами поверхностной структуры) нет взаимно-однозначного соответствия. Состав набора и названия отдельных падежей (ролей) не являются окончательным. В системе TOURUS использованы такие семантические роли: *агент, объект, кого-что, содержание, инструмент, результат, источник, направление, место и нейтральная.* Филлмором предложено несколько падежных систем, охватывающих различные аспекты значений определенных глаголов, одна из которых содержит следующие семантические роли: *агент, контрагент, объект, место, адресат, пациент, результат, инструмент, источник*. Семантические роли Филлмора позволяют учитывать при анализе текста глубинную структуру предложения благодаря соответствующему предварительному описанию модели мира в терминах «ролей». Подход, предложенный Шенком, заключается в представлении идентичных предложений, выраженных различными поверхностными структурами, единой «концептуальной» конструкцией. Шенк исходит из того, что одной из причин разнообразия поверхностных структур при единстве смысла является разнообразие «поверхностных глаголов», описывающих одну и ту же ситуацию. Следовательно, существуют «глубинные» канонические глаголы, унифицирующие в глубинных структурах смысл многих «поверхностных» глаголов. Группой исследователей под руководством Р. Шенка разработана модель концептуальной зависимости для реализации, в частности, следующих

возможностей: 1) получения по тексту на естественном языке его концептуального семантического представления; 2) представления смысла в терминах «атомов» смысла с тем, чтобы вскрывать смысловые сходства и различия между словами; 3) придания синонимичным фразам идентичного семантического представления (СЕМП), а сходным фразам – сходных СЕМП.

Основными семантическими средствами, используемыми в модели концептуальной зависимости (КЗ), являются: 1) знания о языке, хранимые в словаре, и об окружающем мире, хранимые в семантической памяти, выраженные в терминах семантических атомов и сценариев; 2) детальная классификация английских слов; 3) комплекс правил, позволяющих делать умозаключения об обрабатываемом тексте на основе модели знаний.

На основе модели КЗ разработан ряд систем, работающих на сильно ограниченном наборе английских слов.

В наших работах вводится и описывается понятие концептул как *смысловых единиц (обобщенных семантических единиц) семантической структуры текста, отражающих роль лексем и в определенном их сочетании формирующих ожидаемый логически правильный смысл текста в управляемом контексте.*

Главным отличием концептул от семантических ролей Филлмора или семантических единиц модели концептуальной зависимости Шенка является не только их привязка к смысловой ситуации, пусть даже конкретной, а также к структуре значений вопроса, т.е ответных текстов. Иначе говоря, в этом случае можно говорить о разных уровнях конкретизации смысла, о разных глубинах раскрытия смысла. Семантические роли Филлмора и семантические единицы Шенка – это скорее элементы уровня знаний по распознаванию смысла, а концептулы же – элементы уровня знаний по управлению идентификацией ожидаемого смысла. Таким образом, уровень распознавания текста на базе концептул определяется между «поверхностным» (т.е. лексическим) и «глубинным» (т.е. семантическим).

Методы семантического кодирования, независимого от предметной области, как основы универсальной интерпретации, раскрываются, в частности, в работах Лезина Г. В. и др., Поспелова Д. А. и Мартынова В. В. изложена идея, весьма близкая к идее выделения концептул и индивидуальных концептульных грамматик, представляющих собой схемы сочетания концептул («пра-

вильных формул»). Здесь утверждается, что в основе семантического кодирования должен лежать некоторый ограниченный, но достаточно полный набор смысловых атомов (семантических множителей), имеющих характер лингвистических универсалиев, т.е. не зависящих от свойств конкретного ЕЯ и вместе с тем присущих множеству языков. Смысл же сообщения должен передаваться «структурными формулами», в которых смысловые атомы скомбинированы по достаточно жестким правилам. Пример знаковой системы представляет собой попытку оформления концепции кодирования, предложенной В. В. Мартыновым, в виде «языка программирования», т.е. в более технологичном, удобном для реализации виде.

Идеи необходимости привлечения семантики для корректной интерпретации ЕЯ-текстов, и даже необходимости практического приложения семантики широко представлены в публикациях, описываемых в фундаментальной обзорной работе R.Cole. Идеи, близкие к вложенным семантическим классам ответов, выделению концептул и описанию концептуальной грамматики, как это определяется Сулеймановым Д. Ш., высказаны также в работе [107-108]. Как справедливо излагают авторы, большие единицы текста мы понимаем, комбинируя наше понимание через более маленькие. Главная цель лингвистической теории – показать, как эти большие единицы значений складываются (вырастают) из комбинации более маленьких. Это моделируется грамматическими значениями. Далее вычислительная лингвистика пытается реализовать этот процесс эффективным образом. Традиционным является разбиение задачи на синтаксис и семантику, где синтаксис описывает, каким образом различные формальные элементы единиц текста, чаще всего предложение, могут быть комбинированы и семантика описывает, как она интерпретируется. Во многих приложениях языковых технологий кодируемые лингвистические знания, т.е. грамматика, разделены от компонентов обработки. Грамматика состоит из лексикона и правил, которые комбинируют слова и фразы в большие фразы и предложения. Целый ряд языков представления был разработан для кодировки лингвистического знания. Некоторые из этих языков развиты в сторону соответствия с формальной лингвистической теорией, другие развиты для облегчения определенных моделей обработки или специальных приложений. В частности, одним из таких язы-

ков описания семантики текста и является совокупность концептул и их сочетаний – ИКГ. В работе [109] раскрывается понятие *shallow parsing («мелкий» анализ)* – поверхностный («мелкий») грамматический разбор – как обобщенный терм для анализа, менее полный, чем стандартный синтаксический разбор. На выходе «мелкий» анализ не дает дерева непосредственных составляющих. «Мелкий» анализатор может идентифицировать некоторые составляющие фразы, такие как, именные фразы, без выявления их внутренней структуры и их функции в предложении. Другой тип «мелкой» грамматики определяет функциональные роли некоторых слов, таких как главный глагол и его прямые аргументы, что сравнимо с концептулами и ИКГ.

В работе [110] дается обзор сильно продвинутого и распространенного класса лингвистических формализмов – так называемых грамматических формализмов, основанных на правилах (ограничениях) *(constraint-based grammar formalisms)*. Наиболее часто используемые грамматические модели, основанные на правилах – Functional Unification Grammar (FUG) (Kay, 1984), Head-Driven Phrase-Stracture Grammar (HPSG) (Pollard&Sag,1994), Lexical Functional Grammar (LFG) (Bresnan, 1982), Categorial Unification Grammar (CUG) (Kartunen, 1989; Uszkoreit, 1986), and Tree Adjunction Grammar (TAG) (Joshi & Schabes, 1992). Для этих грамматических моделей были разработаны и реализованы мощные формализмы, например, LFG [Bresnan,1982], PATR [Shieber, Uszkoreit, et al., 1983], ALE [Carpenter,1992], STUF [Bouma, Koenig, et al., 1988], ALEP [Alshawi, Arnold, et al., 1991] и др. Существенной составляющей всех этих формализмов являются сложные формальные описания грамматических единиц (слов, фраз, предложений) множеством пар «атрибут-значение», так называемых характеризующих термов. Характеризующие термы могут быть вложенными, т.е. значения могут быть как атомарными символами, так и термами характеристик. Формализмы различаются по разным аспектам. Некоторые из них ограничены для применения к характеристическим термам с простой унификацией. Другие служат для более мощных типов данных, таких как дизъюнктивные термы, функциональные составляющие и т.д. Сила унифицированных грамматических формализмов заключается в их преимуществе для инженерии грамматик. Практика показывает, что большие грамматики хотя, гипотетически, и мо-

гут быть описаны, их реализация чрезвычайно сложна, практически, нереальна. В работе Рахилиной Е. В. описывается весьма интересная и плодотворная идея автора падежной грамматики Ч. Филлмора, первым сформировавшего весьма популярные в компьютерной лингвистике понятия «глубинных падежей»: агенс, пациенс, место и т.д.. Помимо «глубинных падежей», или так называемых «семантических ролей», являющихся ролевыми смысловыми единицами некоторого контекста, Ч. Филлмором предложена идея конструкций (constructions), которая вбирает в себя более широкий и комплексный план. Рассматривается идея таких языковых выражений, у которых есть аспект плана выражения или плана содержания, не выводимый из значения или формы их составных частей**.** Конструкции не сводятся к составляющим и отношениям между ними – в них, кроме того, есть еще значение самой конструкции**,** которое, в частности, накладывает те или иные ограничения на участников конструкции (*кто, что, сколько* участников и т.п.). Конструкции Филлмора очень близки к ИКГ, которые также являются не просто эквивалентными логически правильными схемами текстов, но и определяют специфику этих текстов, классифицируясь на такие типы как ФУНКЦИЯ**,** ДЕТАЛИЗАЦИЯ**,** ПРИЧИНА и т.п. Конструкции же из сегментов, представляющих собой схемы ИКГ, формируют грамматику конструкций – концептуальную грамматику (КГ) как основу прагматически-ориентированной вопросно-ответной диалоговой модели.

Важной проблемой, являющейся частью вопроса формализации высказывания, но в настоящее время не имеющей удовлетворительного решения, является задача сегментации, т.е. разбиение текста на части, порции, элементы, являющиеся теми самыми конструкциями Филлмора, из которых складываются более сложные структуры. В работе [98] рассматриваются различные методы сегментации, в частности, предложен рекурсивный алгоритм сегментации, в работе Liang, Ahmadi, et al. [96] добавлена к этому алгоритму контекстная информация и «спеллчекер» для исправления ошибок, вызванных неправильной сегментацией. Практически во всех проанализированных работах, включая и работы Honda Takeo, Motizuki Hajime, Tu Bao Ho, Okumura Manabu, сегментация текста осуществляется по поверхностным, орфографическим признакам, т.е. знакам пунктуации,

в редких случаях принимаются во внимание лексические признаки, такие как вводные слова. Семантические признаки, учитывающие объектно-предикативные связи, практически, в доступной нам литературе не рассматриваются. Проблемы сегментации вопросно-ответных текстов для рекурсивного приенения правил ИКГ при обработке ответа обучаемого исследуются в ряде работ Сулеймановым Д. Ш.

1.6. ОБЪЕКТНО-ПРЕДИКАТНАЯ СИСТЕМА КАК СОСТАВЛЯЮЩАЯ КОНЦЕПТУАЛЬНО-ФУНКЦИОНАЛЬНОЙ МОДЕЛИ

Ценность и прагматическая ориентированность концептуально-функциональных моделей, отнесенных нами к лингвистическим моделям класса 4, главным образом, заключается в том, что они являются универсальными, наиболее полными описаниями единиц языковых уровней и их грамматик, и являются, как было отмечено выше, «строительным материалом», из которого на основе прагматически-ориентированной технологии строятся лингвистические модели как основа систем обработки ЕЯ-текстов определенного типа. На основе максимально полных концептуально-функциональных моделей для составляющих элементов на разных языковых уровнях мы можем организовать тот самый эффективный асинхронный децентрализованный анализ продукта языковой деятельности, о которой говорится в работах Нариньяни А.С.

Несмотря на методологическую, теоретическую и практическую ценность, такого рода модели крайне слабо исследованы и отражены в литературе. В языкознании единицы языковых уровней, как правило, изучаются и описываются на своем уровне поверхностного проявления: *фонемы* – на уровне фонологии, *морфемы* – на уровне морфологии, *словоформы* – на уровне синтаксиса и т.д., в лучшем случае – в связи с единицами соседних языковых уровней. Модель морфем [62] является уникальным подходом в рамках прагматической технологии, когда морфема исследуется и описывается («инвентаризуется») в структурно-функциональной модели во всех ее проявлениях в языке, на всех языковых уровнях. Главное заключается в том, что сама модель, ее характеристики, уровни и структурные и функциональные связи способствуют лучшему и быстрому раскрытию не всегда

явно выраженных свойств морфем. И наоборот, лингвистическая интуиция высококлассного специалиста, эксперта, «укладывающего» морфемы по «полочкам» модели, позволяет обнаружить недостающие «полочки» – дополнительные параметры, свойства, связи и т.п.

Также косвенным подтверждением важности и необходимости построения такого рода моделей являются аналогичные исследования, проводимые за рубежом. В работе [111] утверждается близкая мысль о том, что интеллектуальная обработка естественного языка в реальных приложениях требует наличия лексикона, который обеспечивает разработчика богатой информацией о морфологических, синтаксических и семантических характеристиках слов, хорошо структурированных, и которые могут быть эффективно применены. Эти цели могут быть достигнуты разработкой инструментария, который облегчает приобретение лексической информации с машиночитаемых словарей и корпусов текстов, также из баз данных и теоретических знаний о слове, предлагаемых в кодированном виде, необходимом для целей NLP *(Natural Language Proccessing)*. Между тем, практически во всех работах речь идет именно об исследованиях и разработках в связи с разработкой лингвопроцессоров, хотя на эти модели можно и нужно смотреть шире. Это не только база для построения обработчиков ЕЯ-текстов, но и богатая база, претендующая на полноту, а также методологический и технологический инструмент для исследований в самом языке. В этих работах, к тому же, как правило, исследуются языковые единицы, их характеристики на уровне слов (полисемия, описание значений, допустимые употребления слова). В настоящее время, практически не имеется описаний на уровне аффиксов. Это и понятно, ибо в английском и русском языках, для которых разработаны наиболее известные лингвопроцессоры, морфология не играет той роли, что в татарском, обладающем богатой морфологией, характеризующейся «формальной элегантностью и естественной сложностью». Наибольшую сложность при создании структурно-функциональной модели морфем вызвала разработка формальной структуры («полки») для отражения в модели значений (т.е. семантики) морфем. Практически, впервые возникла необходимость описания значения татарских морфем на некотором формальном языке «глубинных структур», т.е. на языке семантического уровня.

1.7. АНАЛИЗ СРЕДСТВ ФОРМАЛЬНОГО ОПИСАНИЯ ЗНАЧЕНИЙ

При описании семантики аффиксальных морфем мы исходим из утверждения, что *каждая морфема используется для кодирования того или иного значения в некотором контексте, отражающем некую локальную «модель мира» или здравый смысл «говорящего», отображаемого в «слушающем».* Использование аффиксальных морфем позволяет существенно сократить количество корневых морфем для передачи (кодирования) некоего смысла, то есть служит как элемент, редуцирующий лексическое пространство, необходимое для формирования контекста. Очевидно, минимизация средств адекватного отображения смысла на уровне текста является положительным фактором и при построении эффективных лингвопроцессоров и информационных технологий обработки естественно-языковых текстов.

Локальная модель мира представляет собой формализованное описание некоторого контекста, отражающего объекты и их отношения. Разделение лексем или групп лексем на объекты и отношения является достаточно условной процедурой и зависит от семантических ролей, исполняемых лексемами или группами лексем, отражающими некие значения в определенном контексте.

Примечание 1. Наша задача заключается не в том, чтобы построить семантическую модель, которая позволяет «понять», определить «смысл» некоторого конкретного или обобщенного текста, а в том, чтобы построить концептуальную локальную модель мира, которая может быть соотнесена со значениями конкретных морфем и, соответственно, размечена соответствующими им поверхностными выражениями аффиксальных морфем.

Как известно, в лингвистике разделяются такие понятия, как значение высказывания (или сущности) и его пресуппозиция. Пресуппозиция определяется как предшествующий контекст, предшествующее знание, или как контекст, в котором происходит определение значения сущности. Очевидно, что понимание сущности, прежде всего, обеспечивается именно пресуппозицией. Следовательно, при определении значений аффиксальной морфемы важно рассматривать ее не только как часть словоформы, а также как составляющую контекста, образующего пресуппозицию.

Идея семантических ролей или семантических падежей достаточно активно исследовалась Ч. Филлмором, П. Уинстоном, Ю. Д. Апресяном и др. Известно, что проявление объектов и отношений в тексте можно рассматривать в трех аспектах: синтаксическом (КАК формируется?), семантическом (ЧТО означает?) и прагматическом (ДЛЯ чего? В каких целях?). На наш взгляд, классификация элементов и назначение определенных ролей элементам или классам элементов и у Ч. Филлмора, и у Ю. Д. Апресяна происходит не на семантической, а на прагматической, целевой основе, т.е. по их назначению, а не по семантическому содержательному признаку, как это определяется авторами и в ряде обзорных работ. Поэтому, вместо принятой в литературе понятия «семантическая роль», нами вводится новое понятие – *«прагматическая роль» элемента, означающее целевую функцию объекта в заданном контексте.* Имеется множество работ, в которых с той или иной подробностью приводится и используется объектно-предикатное описание проблемной области, дискурса или контекста. Здесь мы приведем анализ тех объектно-предикатных систем, которые, на наш взгляд, практически, покрывают все пространство объектов и их отношений, описанных во всех упомянутых работах, тем не менее, оставаясь лишь одним из вариантов выделения и описания прагматических ролей, не претендующим на завершенность и даже на достаточность. Подробный анализ ролевых систем Филлмора и Шенка приведен в работе Сулейманова Д. Ш., где также показано, что наиболее полными являются системы отношений, приведенные в работах Закиева М. З., Осипова Г. С. и группы минских ученых.

В работе Закиева М. З. подробно излагается современный взгляд на разновидности семантических компонентов предложения и описывается объектно-предикатная система для татарского языка. Описание модели семантики словосочетаний и богатая информация о значениях аффиксальных морфем, представленная в Академической грамматике татарского языка, является хорошим материалом для заполнения на их основе аспекта «Семантика аффиксов» в модели татарских морфем. Далее приведем крупноблочное описание объектно-предикатной системы Закиева М. З.

Предикаты – отношения, связи (действия или состояния):

1) предикат действия – *мин футбол уйныйм* – ‘*я играю в футбол*’,

2) предикат движения – *туп капкага очып керде – ‘мяч залетел в ворота’,*

3) предикат чувственного восприятия – *мин кояшка шатландым – ‘я обрадовался солнцу’,*

4) предикат речи – *син шигырь сөйлә – ‘ты расскажи стих’,*

5) предикат состояния – *бала йоклый – ‘ребенок спит’,*

6) предикат долженствования – *сиңа түләргә кирәк – ‘тебе надо платить’,*

7) предикат предположения – *ул китте булса кирәк – ‘похоже, он ушел’,*

8) предикат позволения – *сиңа карарга ярый – ‘тебе можно смотреть’*

9) предикат квалификации – *минем сеңелем -җырчы – ‘моя сестра – певица’,*

10) предикат-материал – *күперләре – таштан – ‘у них мост из камня’,*

11) предикат детерминации – *егет бик ьлгер – ‘парень очень умелый’,*

12) предикат принадлежности – *бу китап синеке – ‘эта книга твоя’,*

13) предикат обладания\отсутствия – *ул атсыз калды – ‘он остался без коня’*

14) предикат наличия\отсутствия – *битендә елмаю күренә – ‘на лице видна улыбка’,*

15) предикат предназначения – *бу китап сиңа бирелде – ‘эта книга дана тебе’,*

16) предикат цели – *синең яратуыңны телим – ‘хочу, чтобы ты любила’,*

17) предикат времени – *театрдан иртә кайттык – ‘с театра вернулись рано’,*

18) предикат места – *безнең авыл елга буенда – ‘наша деревня у реки’,*

19) предикат сравнения – *син әниеңә охшаган – ‘ты похожа на свою маму’,*

20) предикат порядковый – *ул беренче килде – ‘он пришел первым’,*

21) предикат количества – *күп болыннар юкка чыккан – ‘многие луга исчезли’.*

субъект – предмет суждения, это то, о чем говорится, о чем сообщается (утверждается, что это спорная категория) – *мин көләм – ‘я смеюсь’.*

объект – это то, на что направлено действие или состояние:

1) объект воздействия – *яңгыр күңелне бозды – ‘дождь испортил настроение’,*

2) объект активного воздействия, или контрагент – *без яшеннән качтык – ‘мы спрятались от грозы’,*

3) объект совместного действия, или коагент – *утынны әти белән кистек – ‘дрова пилили вместе с отцом’,*

4) объект пассивного действия – *бура күтәрелгән – ‘сруб поднят’,*

5) объект содержания речи – *диктор хәбәрләр тапшыра – ‘диктор передает сообщения’,*

6) объект-место – *авылда яши – ‘живет в деревне’,*

7) объект неожиданности – *күлмәккәчә чишенгән – ‘разделся до рубашки’,*

8) объект попутный – *ат белән бергә сарыклар чыгып качты – ‘вместе с лошадью убежали и овцы’,*

9) объект, не ставший им – *чыбыркы урынына алды – ‘взял вместо кнута’,*

10) объект опережающий – *яңгырдан соң кояш чыкты – ‘после дождя выглянуло солнце’,*

11) объект в качестве исполнителя – *ул тегермән булып уйнады – ‘он выступал в качестве мельницы’,*

12) объект-исполнитель – *өй эшләрен энесеннән эшләтте – ‘заставил делать брата домашние дела’.*

В работе Осипова Г. С. выделено и описано 17 видов семантических связей:

1) генеративная связь, один компонент которой обозначает лицо или предмет, принадлежащий некоторой совокупности, категории, обозначаемой вторым компонентом: *сыер – йорт хайваны – ‘корова – домашнее животное’,*

2) дестинативная связь, один компонент которой обозначает назначение для другого компонента*: бу солы атка – ‘этот овес для лошади’,*

3) директивная связь, в которой один компонент обозначает путь, направление другого компонента: *урманга бара – ‘идет в лес’,*

4) инструментальная связь, один компонент которой обозначает орудие действия, обозначаемого другим компонентом: *балтачының балтасы – 'топор плотника',*

5) каузальная связь, один компонент которой обозначает причину проявления другого компонента спустя какое-то время: *чәчкән орлык үсеп чыкты – 'проросло посаженное зерно',*

6) комитативная связь, один компонент которой обозначает сопровождающее другой компонент действие, сопутствующий предмет, сопровождающее лицо: *самолет артыннан эз сузылды – 'за самолетом потянулся след',*

7) коррелятивная связь, один компонент которой выражает возможность наблюдения другого компонента или соответствия предмета другому предмету, компоненту: *күзлек кисәм укый алам – 'в очках могу прочитать',*

8) негативная связь, один компонент которой отрицает, исключает возможность появления другого компонента: *чүп үссә уңыш булмый – если растет сорняк – 'урожая не будет',*

9) лимитативная связь, один компонент которой обозначает сферу применения, назначения другого компонента: *кишер – кимерер өчен – 'морковь – чтобы грызть',*

10) медиативная связь, один компонент которой имеет значение способа, средства действия другого: *аркасында йөзә – 'плывет на спине',*

11) поссесивная связь, один компонент которой выражает отношение владения другим компонентом: *әтинең карандашы – 'карандаш папы',*

12) потенсивная связь, в которой один компонент приводит к увеличению возможности появления другого спустя некоторое время: *ашлама кертсәң тизрәк үсә – 'с удобрением растет быстрее',*

13) результативная связь, в которой один компонент выражает следствие действия второго: *мин куак утырттым – 'я посадил дерево',*

14) репродуктивная связь, в которой один компонент обозначает исходную точку для воспроизведения или превращения для другого компонента: *бәлешне мичтә пешердек – 'пирог испекли в печи',*

15) ситуативная связь, в которой один компонент обозначает ситуацию, определяющую состояние или область действия вто-

рого компонента: *туй авылда булачак – ‘свадьба состоится в деревне’,*

16) трансагрессивная связь, в которой один компонент обозначает результат превращения второго: *утын көлгә әйләнде – ‘дрова превратились в золу’,*

17) финитивная связь, в которой один компонент имеет значение цели, назначения другого: *мин укырыга кердем – ‘я поступил учиться’.*

Под семантической связью в [51] в общем случае также понимается отношение понятий в понятийной системе предметной области, употребляющийся в качестве синонима понятия предикат. Работа группы исследователей из Минска содержит подробную классификацию отношений между понятиями. Ниже проведена систематизация и описание следующих 14 классов отношений, впервые введенных Поспеловым Д. А. и практически охватывающих все отношения, проанализированные нами по работам, касающимся объектно-предикатных систем. Здесь мы описываем классы отношений в структурном виде, дополняя их соответствующими примерами. Как нам представляется, названия классов и подклассов и примеры достаточно мнемонично и доступно отражают суть этих классов и нет необходимости «накручивать» их дополнительными строгими определениями.

1. Отношения классификации. 1.1. Иметь имя («Собаку звали Джек»). 1.2. Класс-подкласс («Органическое соединение – спирт»). 1.3. Часть-целое («Колесо трактора»). 1.4. Элемент-класс («Домашнее животное – корова»). 1.5. Род-вид. («Млекопитающие – парнокопытные»). 1.6. Вышестоящее-нижестоящее («Ректор – декан»). 1.7. Быть эталоном («Победитель олимпиады»).

2. Признаковые отношения. 2.1. Иметь признак («Цвет объекта»). 2.2. Иметь значение признака («Синий»).

3. Количественные отношения. 3.1. Иметь меру («Вес объекта»). 3.2. Иметь значение меры («5 кг»).

4. Отношения сравнения. 4.1. Равно («Все стороны равностороннего треугольника равны»). 4.2. Сравнимо («Вес объекта и вес части объекта»). 4.3. Больше («Индюк больше курицы»). 4.4. Больше или равно («Количество дней в одном месяце больше или равно 28 »). 4.5. Меньше («Плотность льда меньше плотности воды»). 4.6. Меньше или равно («Количество листьев на дереве

меньше или равно количеству почек»). 4.7. Несравнимо («Вес объекта и цвет объекта несравнимы»).

5. Отношения принадлежности («Одежда Марата»).

6. Временные отношения («скорый поезд пришел после товарного»). 6.1. Быть одновременно («Марат и Азат пришли к началу занятий»). 6.2. Быть раньше («До яйца была курица»). 6.3. Быть позже («Яйцо появилось после курицы»). 6.4. Совпадать во времени («Время отлета самолета и отхода поезда в Москву – 19=00»). 6.5. Пересекаться во времени («В три часа обе машины будут проезжать Казань»). 6.6. Быть внутри по времени («Во время твоего пребывания в Казани мы сходим в театр»). 6.7. Начинаться одновременно («Свисток судьи оповестил о начале бега на 5 и 10 тысяч метров»). 6.8. Кончаться одновременно («Мое терпение лопнуло в тот момент, когда заглох мотор»).

7. Пространственные отношения. 7.1. Совпадать в пространстве («И шайба и клюшка оказались в воротах»). 7.2. Быть слева («Слева от дерева стояла машина»). 7.3. Быть справа («Справа от машины зеленело дерево»). 7.4. Быть спереди («Перед преподавателем сидели два студента»). 7.5. Быть сзади («Далеко за горами виднелись облака»). 7.6. Наискосок («Чуть сбоку от дороги вдали светились огни»). 7.7. Пересекаться в пространстве («Над деревом сошлись два облака»). 7.8. Касаться («Облака плывут касаясь крыши домов»). 7.9. Находиться на («Стол стоит на полу»). 7.10. Быть сверху («Перьевые облака плывут выше дождевых»). 7.11. Быть снизу («Подо льдом мирно текла река»). 7.12. Находиться в («В кабине сидело пять человек).

8. Каузальные отношения. 8.1. Быть целью («Мы хотим покорить вершину»). 8.2. Быть мотивом («Он нарушил клятву»). 8.3. Причина-следствие («Горячий уголь прожег материал»).

9. Инструментальные отношения. 9.1. Служить для («Бревно подпирает ворота»). 9.2. Быть средством для («Он доехал до леса на машине»). 9.3. Способствовать («Он предоставил ему свое ружье»). 9.4. Быть инструментом («Обезьяна палкой сшибла банан»). 9.5. Быть вспомогательным средством («У него на поясе висела веревка на случай сильного течения реки»).

10. Информационные отношения. 10.1. Быть отправителем («Он передал письмо для любимой). 10.2. Быть получателем («Мне сегодня пришло письмо»). 10.3. Быть источником информации («Он сообщил мне, что заказ готов»).

11. Порядковые отношения. 11.1. Быть следующим («После Сидоровых пришли Ивановы»). 11.2. Быть очередным («За весной настала очередь лета»). 11.3. Быть ближайшим («Зеленодольск – ближайший к Казани город»).

12. Модальные отношения. 12.1. Возможность («Самолет, который стоит на поляне, полетит к вечеру»). 12.2. Действительность («На фоне заката летит самолет»). 12.3. Необходимость («Для вывоза урожая требуется пять бортовых машин»).

13. Модификаторы («Желательно, чтобы Вы не опоздали к началу сеанса»).

14. Квантификаторы. 14.1. Квантор общности («Все студенты первого курса сдали экзамен по ЭВМ и программированию»). 14.2. Квантор существования («Нашелся студент, который не смог решить квадратное уравнение»).

Как это видно по рассмотренным работам, выделение классов предикатов и объектов есть процесс перманентный, требующий глубокой лингвистической интуиции от автора. Очевидно, ни одна из исследованных классификаций не является полной и завершенной и вряд ли вызовет сомнение у исследователей утверждение, что вопросы полноты и достаточности объектно-предикатной системы могут решаться лишь в ходе практического ее использования, причем, лишь для какой-то фиксированной ситуации. 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.

1.8. АНАЛИЗ СИСТЕМ ОБРАБОТКИ ЕЯ-ТЕКСТОВ НА ОСНОВЕ КОНЦЕПТУАЛЬНО-ФОРМАЛЬНОЙ МОДЕЛИ

Концептуально-формальные модели обеспечивают целевую обработку текстов согласно соответствующим формальным правилам определенного языкового уровня. Ряд работ Сулейманова Д. Ш, Гильмуллина Р. А., Шафигуллина Р. Н. посвящены описанию морфологических моделей татарского языка как концептуально-формальных моделей – двухуровневой автоматной модели, лежащей в основе двухуровневого морфологи-

ческого анализатора, и генеративной модели, лежащей в основе татарского морфологического корректора. В этой связи анализ концептуально-формальных моделей осуществляется на примере работ, содержащих описания различных моделей морфологии.

В последние 10-15 лет наиболее результативно (по сравнению с другими областями обработки языка) в сторону реальных приложений развивалась вычислительная морфология, в которой ключевыми являются: 1) морфологические альтернативы – одни и те же морфемы могут быть реализованы различными путями, в зависимости от контекста; 2) морфотактика: основы, аффиксы и составляющие части не комбинируются свободно, и морфологический анализатор должен уметь определять правильные сочетания морфем. Распространенным для решения первой проблемы является применение метода *cut-and-paste*. Каноническая форма строится путем удаления и присоединения букв к концу слова. Именно такой подход использован нами при построении татарского морфологического корректора, а также программы лемматизации татарских словоформ, включенный в состав инструментального комплекса структурно-функциональной модели морфем.

Технологии конечного состояния для автоматического распознавания и генерации словоформ было введены вначале 1980-х годов. В основе лежит предположение, что правила морфологических альтернатив могут быть реализованы трансдьюсерами конечного состояния. Также известно, что возможные комбинации основ и аффиксов могут быть кодированы как сеть конечного состояния. Первая практическая система, включающие эти идеи – двухуровневая модель для татарского языка, подробно представлена в работах Сулейманова Д. Ш., Гильмуллина Р. А. Система основана на множестве деревьев связанных букв для лексикона и параллельных трансдьюсеров конечного состояния, кодирующих морфологические альтернативы. Двухуровневый распознаватель отображает поверхностную строку в последовательность веток в дереве букв, используя трансдьюсеры, и вычисляет основу исходя из информации, имеющейся в границах веток. В разработках, связанных с этими, было замечено, что огромные списки слов, проверяемых на правильность, могут быть скомпилированы в удивительно компактный автомат конечных состояний (Appel & Jacobson, 1988; Lucchesi & Kowaltowski, 1993). На таком под-

ходе – использовании двухуровневой морфологической модели, устроен татарский морфологический анализатор, реализованный в среде программного инструментария PC-KIMMO.

К концептуально-формальным относятся, в частности, модели автоматизированных корректоров, осуществляющих проверку орфографии в ЕЯ-текстах. Практически большинство известных моделей для русского языка представляют собой компьютерную реализацию словаря Зализняка. Модели корректоров различаются интерфейсом и дополнительными функциями обработки сопутствующей информации, принципами реализации (резидентно – ОРФО, ОРК-Т, RUSC; в виде встроенной модули – спеллчекер: ОРФО, ТАТЕДКОР, Пропись; или в виде отдельной программы – ТАТКОР, Грамота), но проверка корректности текста, как правило, реализуется по схожему алгоритму следующим образом. На основе исходного словаря основ порождается список словоформ, далее этот список некоторым образом упаковывается и в нем реализуется функция поиска. Таким образом, проверка правильности написания слова сводится к поиску в словаре соответствующей словоформы. При этом процесс порождения словоформ, достаточно сложный для русского языка, осуществляется вне этапа непосредственной проверки. По этому принципу возможно построение корректора для любого языка: берется большой объем текстов и на его основе формируется список словоформ (так, в частности, создаются корректоры в системе оптического распознавания FineReader). Такой корректор может допускать ошибки, но используя его, проще проверять тексты, содержащие специальные термины, и создание словаря для нового языка не занимает много времени. Именно используя такой подход (так называемый, парадигматический) был построен один из первых коммерческих татарских корректоров ОРК-Т группой «Стагирит» из г.Екатеринбурга под руководством Ахметьянова Р. Парадигматический способ организации проверки правильности словоформы, наиболее широко используемый в корректорах, лежит в основе, в частности, также и грамматического и стилистического корректора ОРФО. Первыми пакетами программ, возможности которых были проанализированы и учтены при разработке первого татарского генеративного корректора были корректоры RUSP, ОРФО, ОРК-Т (для русского и татарского языков), описание которых дается в работе автора.

Однако эта модель, успешно работая на относительно большом пространстве текста, не покрывает всевозможные правильные словоформы в силу специфики татарского языка. При анализе татарских текстов установлено, что хотя в 90% случаях к основе присоединяется не более 3-4 аффиксов, в некоторых случаях (не более 3%) возможно участие гораздо большего числа аффиксов (более 10) в образовании татарской словоформы, к тому же, число порождаемых словоформ от одной основы потенциально неограниченно. В работе Ашманова И.С. содержатся классификация существующих систем проверки правописания, описание стандартных наборов функций, анализ принципов построения существующих грамматических и стилистических корректоров и типов обнаруживаемых ими ошибок.

Выводы

В результате анализа работ в области разработки систем и технологий автоматизированной обработки знаний получен ряд выводов, позволивший определить направление, структуру и содержание исследований:

1. 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, и ряд других идей, весьма четко и убедительно очерчивают перспективные направления развития систем обработки информации, в том числе ЕЯ-текстов.

В ряде 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.*

2. Отсутствие универсальных моделей обработки ЕЯ-текстов является *косвенным подтверждением правильности прагматически-ориентированного подхода,* когда исследуются и находятся такие проблемные области, для которых удается построить вполне удовлетворительные лингвистические модели, т.е. объединить в рамках единой технологии концептуально-формальные

средства, достаточные с методологической и практической точек зрений. Очевидно, что если мы начнем строить лингвистические процессоры только при полном наличии моделей всех языковых уровней, скорее всего, никогда не будет достигнут этап серьезной практической реализации. Построение реальных систем обработки информации на основе прагматически-ориентированной технологии является хорошим компромиссным вариантом.

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

3. Выделение классов предикатов и объектов в «модели мира» есть процесс перманентный, требующий глубокой лингвистической интуиции от автора. Ни одна из рассмотренных классификаций предикатов и объектов не является полной и завершенной и, очевидно, что вопросы полноты и достаточности объектно-предикатной системы могут решаться лишь в ходе практического ее использования. Следовательно, весьма *актуально иметь некий инструментарий, включающий совокупность семантических универсалий, 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.

4. 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 и др.), учет которых позволяет строить прагматически-ориентированные лингвистические модели как основы эффективных анализаторов ответов обучаемого.*

5. *Сила унифицированных грамматических формализмов, как основы семантического анализа ЕЯ-текстов, в преимуществе их возможностей для разработки прагматически-ориентирован-*

ных 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.

6. 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.*

7. Большинство моделей корректоров реализует *парадигматический* подход к определению правильности лексемы, как наиболее подходящий для группы индо-европейских языков. Другим подходом, распространенным для решения проблемы морфологических альтернатив (определение корректности соответствующего аффикса) является также применение метода *cut-and-paste,* заключающегося в определении правильной формы путем удаления и присоединения букв к концу слова. Именно такой *подход, являющийся частью генеративного алгоритма, использован нами при построении татарского морфологического корректора.* Данная модель морфологии используется при распознавании татарских текстов в известной системе OCR FineReader фирмы ABBYY. Третий подход, используемый для разработки концептуально-формальных морфологических моделей – *использование технологии конечного состояния для автоматического распознавания и генерации словоформ.* На таком подходе – использовании двухуровневой морфологической модели, устроен татарский морфологический анализатор, реализованный в среде программного инструментария PC-KIMMO. Двухуровневый морфологический анализатор, благодаря лучшим временным характеристикам, в настоящее время наиболее активно используется на практике. В частности, именно такой анализатор внедрен в состав модуля поиска УИС «Россия» (МГУ), используется как программный инструмент для аннотации текстов в национальном корпусе татарского языка «Татар теле».

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Абрамов В. Г., Брябрин В. М., Пховелишвили М. Г. ДИЛОС – диалоговая система для взаимодействия на естественном языке. Сообщения по программному обеспечению ЭВМ. – М.: 1979. – 41 с.

2.Абросимов А., Гатауллин Г. С., Сулейманов Д. Ш. и др. Технологический комплекс обучения КГУ: методология обучения с использованием ЭВМ // *Рукопись деп. в ВИНИТИ N 916-87*. – Казань: 1987. – 43 с.

3. Абросимов А., Гатауллин Г. С., Исмагилов Н. А., Сулейманов Д. Ш. и др. Технологический комплекс обучения КГУ: программно-методическое обеспечение учебного процесса // *Рукопись деп.в НИИ Проблем ВШ N 1324-87*. – Казань: 1987. – 37 с.

4. Андрусенко Т. Б. Лингвистические структуры в компьютерных учебных средах. – Киев: Наукова Думка, 1994. – 160 с.

5. Анисимович К., Селегей В. О роли лингвистических технологий в оптическом распознавании полиязычных текстов // *Труды международного семинара Диалог-96 по компьютерной лингвистике и ее приложениям под редакцией А. С. Нариньяни*. – 1996. – С. 28–30.

6. Апресян Ю. Д. Образ человека по данным языка: попытка системного описания // *Вопросы языкознания*. – 1995. – № 1. – С. 37–67.

7. Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л. и др. Лингвистический процессор для сложных информационных систем. – М.: Наука,1992. – 256 с.

8. Ахоу А. В. Индексные грамматики – расширение контекстно-свободных грамматик // *Кибернетический сборник*. – Вып.5. – М.: 1962. – С. 15–21.

9. Ашманов И. С. Архитектура и технология промышленной реализации прикладных лингвистических систем (проверка правописания и электронные словари) // *Автореферат дисс. на соискание ученой степени кандидата технических наук*. – Переславль-Залесский: 1995. – 22 с.

10. Бейлин Дж. Краткая история генеративной грамматики // *Сб. «Фундаментальные направления современной американской лингвистики»* / Под ред. А. А. Кибрика, И. М. Кобозевой, И. А. Секериной. – М.: Изд-во МГУ, 1997. – С. 13–57.

11. Бельнап Н., Стил Т. Логика вопросов и ответов. Пер. с англ. Г. Е. Крейдлина. – М.: Прогресс, 1981. – 288 с.

12. Братчиков И. Л. Экспертные системы и проблема анализа ответов обучаемых // *Бюллетень «Экспертные системы и ПРОЛОГ в учебном процессе II». Краткое изложение докладов на школе-семинаре*. – Йошкар-Ола: ОП СНИО, 1990. – С. 18–24.

13. Брусиловский П. Л. Интеллектуальные обучающие системы // *Информатика. Научно-технический сборник. Серия «Информационные технологии. Средства и системы»*. – 1990. – Вып. 2. – С. 3–22.

14. ДИЛОС – диалоговая система для взаимодействия с ЭВМ на естественном языке / В. Г. Абрамов, В. М. Брябрин, М. Г. Пховелишвили и др. – Москва: ВЦ АН СССР, 1979. – 80 с.

15. Брябрин В. М., Сенин Г. В. Анализ естественного языка в ограниченном контексте // *Вопросы кибернетики*. – 1980. – Вып. 61. – С. 111–117.

16. Бухараев Р. Г., Сулейманов Д. Ш. Об одном подходе к разработке интеллектуальных АОС // *Кибернетика*. – 1986. – № 3. – С. 42–49.

17. Бухараев Р. Г., Сулейманов Д. Ш. Семантический анализ в вопросно-ответных системах. – Казань: Изд-во Казан. ун-та, 1990. – 124 с.

18. Вакуленко В. Н. База знаний консультирующей системы // *Тез. конф. Восток-Запад по новым информационным технологиям в образовании*. – М.: МЦНТИ, 1992. – С. 37–38.

19. Васильев А. В., Васильева Н. Л. Семантический анализ сообщений обучаемых // *Тезисы докладов межзональной научно-методической конференции*. – Минск: Изд-во БГУ, 1984. – С. 10–11.

20. Васильева Н. Л. Конструирование программ семантического анализа ответа в автоматизированных учебных курсах // *Методы и системы технической диагностики. Научно-техническое направление «Искусственный интеллект в автоматизированных обучающих системах». Межвуз. науч. сб*. – Вып. 8. – Саратов: Изд-во Сарат. ун-та, 1987. – С. 78–79.

21. Виноград Т. Программа, понимающая естественный язык. – М.: 1976. – 283 с.

22. Вудс В. А. Сетевое грамматики для анализа естественного языка // *Кибернетический сборник*. – Вып. 13. – М.: 1976. – С. 121–158.

23. Гвида Дж., Сомальвико М. Общение с системами искусственного интеллекта на естественном языке: проект DONAU // *Сб. научных трудов «Лингвистические процессоры и представление знаний»*. – Новосибирск: 1981. – С. 41–65.

24. Гвида Дж., Тассо К. Алгоритмы и эвристики в системе понимания ЕЯ // *Сб. научн. трудов под ред. А. С. Нариньяни «Прикладные и экспериментальные лингвистические процессоры»*. – Новосибирск: 1982. – С. 37–59.

25. Городецкий Б. Ю. Компьютерная лингвистика: моделирование языкового общения // *Новое в зарубежной литературе*. *Компьютерная лингвистика. Под ред. Городецкого Б. Ю*. – Вып. XXIV. – М.: Прогресс, 1989. – С. 5–31.

26. Гофен А. М., Левин Н. А., Анацкий Н. И., Мясник А. А. Принципы построения обучающих и контролирующих систем нового поколения // *Разработка и применение программных средств ПЭВМ в учебном процессе. Материалы VI Всесоюзного семинара*. – Кн.1. – М.: ИПИ АН, 1991. – С. 122–124.

27. Денинг В., Эссиг Г., Маас С. Диалоговые системы «Человек-ЭВМ». Адаптация к требованиям пользователя / Пер. Котова Ю. Б., под ред. Мартынюка В. В. – М.: Мир, 1984. – 112 с.

28. Довгялло А. М. Диалог пользователя и ЭВМ. Основы проектирования и реализации. – Киев: 1981. – 232 с.

29. Дракин В. И., Попов Э. В., Преображенский А. Б. Общение конечных пользователей с системами обработки данных. – М.: Радио и связь, 1988. – 288 с.

30. Ефимов Е. И. Сфинкс – вычислительный комплекс, предназначенный для обоснования интеллектуальных решений. – М.: ВЦ РАН, 1993. – 20 с.

31. Зайцева Г.В. Алгоритм анализа ответов на базе фреймовой модели // *Эффективность применения автоматизированных обучающих систем в учебном процессе Высшей школы. Тезисы докладов Всесоюз. научно-методического совещания*. – Рига: Изд-во РПИ, 1988. – С. 190–191.

32. Закиев М. З. Татарская грамматика. Т3. Синтаксис. – Казань: Таткнигоиздат, 1992. – 488 с.

33. Зализняк А. А. Грамматический словарь русского языка. Словоизменение. – М.: Русский язык, 1980. – 880 с.

34. Кузин Л. Т., Преображенский А. Б., Хорошевский В. Ф. и др. Разработка и исследование методов построения интеллектуальных вопросно-ответных систем (проект МИВОС). Отчет о НИР. – М.: МИФИ, 1977. – 291 с.

35. Лингвистический процессор для сложных информационных систем / Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин и др. – М.: Наука, 1992. – 256 с.

36. Майлопулос Д. и др. TOURUS – система для управления данными, понимающая естественный язык // *Труды IV Международной объединенной конференции по искусственному интеллекту*. – Т. 2. – М.: 1975. – С. 42–62.

37. Мальковский М.Г. Диалог с системой искусственного интеллекта. – М.: Изд-во МГУ, 1985. – 214 с.

38. Мартынов В.В. Универсальный семантический код: УСК-3. – Минск: Наука и техника, 1984. – 131 с.

39. Мельчук И. А. Опыт теории лингвистических моделей “Смысл-Текст”. – М.: Наука, 1974. – 314 с.

40. Микулич Л. И. Специализирование диалоговых систем // *Сб. «Вопросы обработки прикладных систем»*. – Новосибирск: Изд. ВЦ СО АН СССР, 1979. – С. 111–129.

41. Моделирование языковой деятельности в интеллектуальных системах / Под ред. А. Е. Кибрика, А. С. Нариньяни. – М.: Наука. Главная ред. физ.-мат.лит., 1987. – 280 с.

42. Труды Международного семинара Диалог-96: компьютерная лингвистика и ее приложения (Пущино, 4–9 мая, 1996г.). Под ред. А. С. Нариньяни. – М.: 1996. – 305 с.

43. Нариньяни А.С. Лингвистические процессоры ЗАПСИБ (Часть 1 – Задачи проекта) // *Препринт АН СССР. Сиб. Отд-е. ВЦ: № 199*. – Новосибирск: 1979. – 22 с.

44. Нариньяни А. С. Модель или алгоритм: новая парадигма информационной технологии // *Информационные технологии*. – 1997. – С. 11–16.

45. Невзорова О. А. Машинное обучение и задачи обработки естественного языка // *Новости ИИ*.– 1998. – № 1. – С. 5–23.

46. Нильсон Н. Принципы искусственного интеллекта. – М.: Радио и связь, 1985. – 376 с.

47. Новицкий Л. П. Распознование ответов обучаемых в системе "Контакт" // *Тезисы докладов межзональной научно-методической конференции*. – Минск: Изд-во БГУ, 1984. – С. 12–13.

48. Новое в зарубежной лингвистике. Вып. XXIII. Когнитивные аспекты языка: Пер. с англ. – М.: Прогресс, 1988. – 320 с.

49. Обработка текста и когнитивные технологии / Под ред. А. Г. Дьячко. – Вып. 1. – Москва, Пущино: ОНТИ ПНЦ РАН, 1997. – 116 с.

50. Олдерс Д. А. Генерация и анализ конструкций языков, задаваемых формальными грамматиками в автоматизированных обучающих системах // *Научная организация учебного процесса и применение автоматизированных управляющих и обучающих систем в ВУЗах Лат. ССР. Сб. результатов научно-исследовательских работ по проблемам высшей шк*. – Рига: Изд-во РПИ, 1982. – С. 209–213.

51. Осипов Г.С. Построение моделей предметных областей. Ч. 1. Неоднородные семантические сети // *Известия РАН, сер. «Техническая кибернетика»*. – 1990. – №5. – С. 32–45

52. Осипов Г. С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. – М.: Наука. Физматлит, 1997. – 112 с.

53. Падучева Е. В. Вопросительные местоимения и семантика вопроса // *Разработка формальной модели естественного языка*. – Новосибирск: 1981. – С. 80–105.

54. Поляков В. Н. Проблемы представления, приобретения и использования знаний в свете обработки естественного языка // *Труды Казанской школы по компьютерной и когнитивной лингвистике. Центр им. Н.И. Лобаческого*. – 1999. – Т. 4. – С. 93–110.

55. Попов Э. В. Общение с ЭВМ на естественном языке. – М.: Наука. Главная редакция физико-математической литературы, 1982. – 360 с.

56. Попов Э. В. Принципы построения системы общения пользователя с базами данных // *Лингвистические процессоры и представление знаний*. – Новосибирск: 1981. – С. 66–78.

57. Попов Э. В. Экспертные системы: Решение неформализованных задач в диалоге с ЭВМ. – М.: Наука. Гл. ред. физ.-мат. лит., 1987. – 288 с.

58. Поспелов Д. А. Ситуационное управление: теория и практика. – М.: Наука, 1986. – 284 с.

59. Сборник научных трудов в трех томах. 5-я национальная конференция с Международным участием «Искусственный интеллект-96». – Казань: 1996. – Т. 1 “Компьютерная лингвистика”. – С. 93–158.

60. Селиванова Н. В., Хорошевский В. Ф., Рыбина М. В. Разработка лингвистического процессора одной интеллектуальной информационно-управляющей системы на базе языка ATNL // *Лингвистические процессоры и представление знаний. Тезисы докладов*. – Новосибирск: 1981. – С. 79–94.

61. Сокулина И. Н. Структура и задачи АОС интеллектуального типа // *Методы и системы технической диагностики. Научно-техническое направление «Искусственный интеллект в автоматизированных обучающих системах». Межвуз. науч. сб.* – Саратов: Изд-во Сарат. ун-та, 1987. – Вып. 8. – С. 5–7.

62. Сулейманов Д. Ш., Гатиатуллин А. Р. Структурно-функциональная компьютерная модель татарских морфем. – Казань, Изд-во “Фэн”, 2003. – 220 с.

63. Сулейманов Д. Ш., Шафигуллин Р. Н. Морфологический корректор татарских текстов – ТАТКОР // *Сб. трудов школы по компьютерной и когнитивной лингвистике*. – Казань: 2002. – С. 253–255.

64. Сулейманов Д. Ш. Принципы семантической классификации текстов и их анализ по классам в АОС РС // *Математические основы и программное обеспечение автоматизации интеллектуальной деятельности. Матер. третьей научн. конф. молодых ученых и специалистов фак. выч. мат. и киберн. Сб. деп. в ВИНИТИ N 331-84*. – Казань: КГУ, 1984. – С. 53–56.

65. Сулейманов Д. Ш., Гатиатуллин А. Р. Структурно-функциональная компьютерная модель татарских морфем. – Казань: Изд-во “Фэн”, 2003. – 220 с.

66. Сулейманов Д. Ш., Гильмуллин А. А., Гильмуллин Р. А. Двухуровневое описание морфологии татарского языка // *Тезисы Международной научной конференции, посвященной 200-летию университета «Языковая семантика и образ мира»*. – Казань: Изд-во КГУ, 1997. – Кн. 2. – С. 65–67.

67. Сулейманов Д. Ш., Шафигуллин Р. Н. Морфологический корректор татарских текстов ТАТКОР // *Татарский язык и новые информационные технологии. Серия «Интеллект. Язык. Компьютер»*. – Казань: Изд-во Казан. ун-та, 1995. – Вып. 2. – С. 86–90.

68. Сулейманов Д. Ш. Обработка ЕЯ-текстов на основе прагматически-ориентированных лингвистических моделей // *Сб. под ред.*

В. Д. Соловьева «Обработка текста и когнитивные технологии». – 1998. – Вып. 3. – С.205–212.

69. Труды Международного семинара Диалог-95: компьютерная лингвистика и ее приложения / Под ред. Р. Г. Бухараева, А. С. Нариньяни, В. Д. Соловьева. – Казань: 1995. – 362 с.

70. Труды Международного семинара Диалог-96: компьютерная лингвистика и ее приложения / Под ред. А. С. Нариньяни. – М.: 1996. – 305 с.

71. Труды Международного семинара Диалог-97: компьютерная лингвистика и ее приложения / Под ред. А. С. Нариньяни. – М.: 1997. – 315 с.

72. Труды Международного семинара Диалог-98: компьютерная лингвистика и ее приложения. В двух томах / Под ред. А. С. Нариньяни. – Казань.: 1998. – 885 с.

73. Труды Шестой национальной конференции по искусственному интеллекту с Международным участием КИИ-98 в трех томах. – М.: 1998. – 726 с.

74. Уилкс Ч. Анализ предложений английского языка. Новое в зарубежной лингвистике. – М.: 1982. – Вып. XII. – Ч2. – С. 55–69.

75. Уинстон П. Искусственный интеллект. – М.: 1980. – 580 с.

76. Филлмор Ч. Дело о падеже // *Кн. «Новое в зарубежной лингвистике»*. – Вып. X. Лингвистическая семантика. – М.: Прогресс, 1981.

77. Формальное описание структуры естественного языка. – Новосибирск: 1980. – 174 с.

78. Харин Н.П. INTER-SPELL – многоязычная программа автоматической проверки правописания текстов. Документация. Пакет программ.

79. Хомский Н. Три модели описания языка // *Кибернетический сборник*. – 1961. – Вып. 2. – С. 81-92.

80. Хомский Н. Синтаксические структуры // *Сб. «Новое в лингвистике»*. – Вып.2. – Москва: Изд-во ин. лит., 1962.

81. Шаров С. А. Средства компьютерного представления лингвистической информации. Обзор. – [Электрон. ресурс]. – 2006. – URL: http://nl-web/ (дата обращения: 20.02.2016).

82. Шенк Р. Обработка концептуальной информации. – М.: Энергия,1980. – 361 с.

83. Эрли Д. Эффективный алгоритм анализа контекстно-свободных языков // *Языки и автоматы*. – М.: Мир, 1975. – С. 32–42.

84. Antworth E.L. PC-KIMMO: a two-level processor for morphological analysis // *Technical Report Occasional Publications in Academic Computing, Summer Institute of Linguistics*. –Dallas, Texas: 1994. – № 16.

85. Appel A.W., Jacobson G.J. The world's fastest scrabble program // *Communications of the ACM*. – 1988. – Vol.31, № 5. – P. 572–578.

86. Borota J., Hajic J., Hajicova E. etc. TIBAQ (Text-and-Inference Based Answering of Questions) – Text-and-Inference Based Approuch to Question Answering // *Theoretical and Computational Linguistics. V.3. Edited by Eva Hajicova*. – Praha: 1995. – 141 p.

87. Bouma G., Koenig E., Uszkoreit H. A flexible graph-unification formalism and its application to natural-language processing // *IBM Journal of Research and Development*. – 1988.

88. Bresnan J., editor. The Mental Representation of Grammatical Relations. – Cambridge, Massachusetts: MIT Press, 1982.

89. Briscoe E. J. Lexical issues in natural language processing // *Klein E., Veltman F. Natural Language and Speech*. – Springer-Verlag, 1992. – P. 39–68.

90. Cole R.A., Mariani J., Uszkoreit H., et al (editors). Survey of the State of the Art Human Language Technology. – [Электрон. ресурс]. – 1995. – URL: ftp://speech.cse.ogi.edu/pub/docs/HLT/ (дата обращения: 20.02.2016).

91. DeJong G. Prediction and Substantiation: A New Approach to Natural Language Processing // *Cognitive Scince*. – 1979. – Vol. 3 – P. 251–273.

92. Guida G. Ideas About Design of Natural Language Interfaces to Query Systems // *Proc. Workshop on Natural Language for Interaction with Data Bases, IIASA CP-78-9*. – Laxenburg, Austria: 1978. – P. 265–279.

93. Hasling D.W., Clancey W.J., Rennels G. Strategic explanations for diagnostic consultation system // *International Journal on the Man-Machine Studies*. – 1984. – Vol. 20. – № 1. – P. 3–19.

94. Honda Takeo, Motizuki Hajime, Tu Bao Ho, Okumura Manabu // *Poster Papers 9th European Conference On Machine Learning*. – Prague: 1997. – P. 68–77.

95. InterBASE – система конструирования ЕЯ-интерфейсов к базам данных. Вер. 1.3. Документация. Пакет прогр. – Москва-Новосибирск: 1992. – 163 с.

96. Liang S., Ahmadi M., Shidhar M. Segmentation of Touching Characters in Printed Document Recognition // *Proceedings of the Second International Conference on Document analysis and Recognition*. – Tsukuba Science City, Japan: AIPR-IEEE, IAPR, 1993. – P. 569-572.

97. Lucchesi C.L., Kowaltowski T. Applications of finite automata representing large vocabularies // *Software-Practice and Experience*. – 1993. – Vol. 23, № 1. – P. 15–30.

98. Nadle, M. A Survay of Document Segmentation and Coding Tecniques // *Computer Vision and Image Processing*. – 1984. – Vol. 28. – P. 240–262.

99. Neidle C. Lexical Functional Grammar // *Proceedings of the ESSLLI*. – Prague: 1996. – P. 107–128.

100. Netter Klaus. Constraint-Based Grammar Models // *Proceedings of the Eigh Eurepean Summer School in Logic, Language and Information ESSLLI-96*. – Prague: 1996. – P. 1–10.

101. Oflazer Kemal. Two-level Description of Turkish Morphology // *Literary and Linguistic Computing*. – 1994. – Vol. 9, No. 2.

102. Pollard C., Sag I.A. Head-Driven Phrase-Stracture Grammar // *Center for the Study of Language and Information (CSLI) Lecture Notes*. – Chicago: Stanford University Press, University of Chicago Press, 1994.

103. Proceedings of the NeMLap-2. The 2d Int. Conf. on New Methods in Language Processing. – Ankara, Turkey: 1996.

104. Russian Language Processor Russicon: Design and applications // *Proceedings East-West Conference on Artificial Intelligence From Theory to Practice. EWAIC' 93. Organized by RAAI Jointly with ICSTI*. – M.: 1993. – P. 175–180.

105. Shieber S. M., Uszkoreit H., Robinson J., Tyson M. The formalism and Implementation of PATR-II. SRI International. – Menlo Park, California: 1983.

106. Suleymanov D. Sh. Towards a difinition of Tatar cases via the structural-functional morphemic model // *Proceedings of the LP'96 Conference on Languages Typology*. – Prague: 1996. – P. 388–395.

107. Uszkoreit H. Categorial unification grammars // *Proceedings of the 11th International Conference on Computational Linguistics*. – Bonn: ACL, 1986.

108. Zaenen Annie, Uszkoreyt Hans: Language Analysis and Understanding // Survey of the State of the Art in Human Language Technology. – [Электрон. ресурс]. – 1997. – URL: http://www.dfki.de/~hansu/HLT-Survey.pdf (дата обращения: 20.02.2016).

109. Karlsson F., Karttunen L. Sub-Sentential Processing // Survey of the State of the Art in Human Language Technology. – [Электрон. ресурс]. – 1997. – URL: http://www.dfki.de/~hansu/HLT-Survey.pdf (дата обращения: 20.02.2016).

110. Uszkoreita H. & Annie Zaenenn A. Grammar Formalisms // Survey of the State of the Art in Human Language Technology. – [Электрон. ресурс]. – 1997. – URL: http://www.dfki.de/~hansu/HLT-Survey.pdf (дата обращения: 20.02.2016).

111. Sanfilippo A. Lexicons for Constraint-Based Grammars // Survey of the State of the Art in Human Language Technology. – [Электрон. ресурс]. – 1997. – URL: http://www.dfki.de/~hansu/HLT-Survey.pdf (дата обращения: 20.02.2016).

Глава 2. ОНТОЛОГО-ЛИНГВИСТИЧЕСКИЕ СИСТЕМЫ: БАЗОВЫЕ МОДЕЛИ

О. А. Невзорова

2.1. БАЗОВЫЕ ПРЕДПОСЫЛКИ ПРОЕКТИРОВАНИЯ ОНТОЛОГО-ЛИНГВИСТИЧЕСКИХ СИСТЕМ

Повсеместная компьютеризация общества, становление и развитие сетевых информационных технологий обеспечили переход общества в новое качественное состояние глобальной информатизации, что способствовало ускоренному развитию информационных технологий обработки текстовой информации. Основная часть информационных ресурсов общества представлена в естественно-языковой (ЕЯ) форме в виде текстов на различных национальных языках. Поэтому актуальными и востребованными являются технологии информационного поиска, извлечения знаний из текстов, автоматического реферирования, машинного перевода и др.

В настоящее время исследования и разработки в области создания систем IE (Information Extraction) активно ведутся во всем мире [1, 2]. Под извлечением информации (IE) понимается идентификация и семантическая классификация знаний, извлеченных из неструктурированных источников (например, текстов) для задач информационных систем. В последние годы задача IE интегрируется в более крупные приложения, связанные с выбором (поиском) информации для различных целевых задач, в том числе для задач принятия решений. Однако современные системы извлечения знаний из текстов обладают достаточно ограниченными функциональными возможностями, связанными с использованием узкоспециализированных шаблонов для извлечения знаний. Задачи IE-систем тесно пересекаются с классическими задачами ЕЯ-систем (традиционно называемыми лингвистическими процессорами). Современные лингвистические процессоры [3, 4, 42] ориентированы на решение задач морфологического и синтак-

сического анализа естественного языка и используют в качестве формальных моделей различные лингвистические теории синтаксиса [7-11].

Другим перспективным направлением, существенно влияющим на прогресс в области информационных технологий обработки текста, являются исследования, связанные с моделированием онтологической семантики. В этой области активно формируется новая глобальная парадигма Semantic Web [16, 17], в основе которой лежит идея семантического структурирования информационного пространства знаний. Важнейшая роль семантических знаний всегда подчеркивалась в когнитивных исследованиях, таким образом, можно утверждать, что семантический уровень является интегральным системообразующим уровнем языковой системы. Центральной проблемой построения семантического пространства знаний является недостаточная проработанность вопросов формализации семантики ЕЯ. Тем не менее, для эффективного моделирования лингвистических задач (в том числе классических задач морфологии и синтаксиса) требуется привлечение семантических знаний и технологий их использования. Семантика является каркасом языкового интеллекта, который обеспечивает межуровневые связи, разрешение многозначности, связность и компрессию входных и выходных текстов. Анализ ЕЯ-текстов интеллектуальной системой должен осуществляться в широком контексте знаний о внешнем мире (внешняя семантика), с учетом целей и функциональных возможностей системы (внутренняя семантика).

В работе предложен онтолого-лингвистический подход к построению интеллектуальных систем обработки естественного языка, а также методы и технологии обработки ЕЯ-текстов, разработанные в рамках указанного подхода. Суть онтолого-лингвистического подхода к обработке естественного языка заключается в конструктивной интеграции семантических и лингвистических моделей и ресурсов, построении эффективного взаимодействия указанных составляющих при решении прикладных задач. Класс интеллектуальных систем, реализованных в рамках развиваемого подхода, будем называть в дальнейшем онтолого-лингвистическими системами. Центральным ядром онтолого-лингвистических систем является система онтологических моделей, под управлением которой выполняется построение решения

прикладной задачи. В качестве экспериментальной предметной области для онтолого-лингвистических систем выбрана область компьютерного анализа ЕЯ-текстов.

2.2. КОНЦЕПТУАЛЬНАЯ АРХИТЕКТУРА ОНТОЛОГО-ЛИНГВИСТИЧЕСКИХ СИСТЕМ

Онтолого-лингвистические системы ориентированы на решение сложных задач обработки текстов, требующих семантических знаний. Целью создания онтолого-лингвистических систем является обеспечение решения сложных задач обработки текстов путем организации системы взаимодействий различных уровней обработки текста, включая онтологический уровень, связанный с построением семантически адекватной модели предметной области и различных лингвистических уровней, связанных с лингвистическими свойствами и отношениями объектов предметной области. Класс онтолого-лингвистических систем отличается объединением онтологических (экстралингвистических) и лингвистических знаний, эвристических и формальных методов обработки текстов.

В структуре онтолого-лингвистической системы можно выделить две основные взаимодействующие компоненты: онтологическую и лингвистическую. Типовой набор функциональных компонентов онтолого-лингвистической системы представлен на рисунке 1.

Онтологическая компонента поддерживает проектирование системы онтологических моделей и обеспечивает взаимодействие с лингвистической компонентой при разработке лингвистических приложений. Разработка прикладных онтологий может опираться на существующие стандарты разработки онтологий и тезаурусных систем, а также иметь специфические методы.

Онтологическая подсистема обеспечивает поддержку решения следующих задач обработки текстов:

1. Онтологическая разметка текстов элементами (концептами, отношениями) прикладной онтологии. Задача онтологической разметки текста связана с автоматическим распознаванием текстовых описаний онтологических концептов и выделения границ текстовых описаний онтологических составляющих. Этап онтологической разметки текста позволяет выделять онтологи-

ческие компоненты, из которых впоследствии собираются более сложные семантические структуры.

2. Извлечение информации из текстов (распознавание и интерпретация прикладных задач). Классические задачи IE ориентированы на идентификацию в тексте элементов определенных семантических классов, например сведения о персоне, организации и т.п. В общем случае, можно определить классы прикладных задач и онтологические структуры, связанные с прикладной задачей, распознавание и идентификация которых будет являться задачей онтологического анализа.

3. Онтологическая поддержка задач лингвистического анализа. Можно выделить группу задач лингвистического анализа, решения которых могут быть получены на основе механизмов взаимодействия онтологической и лингвистической компонент онтолого-лингвистической системы. Фактически, указанное взаимодействие обеспечивает результаты семантико-синтаксического анализа текстов. Для решения каждой задачи лингвистического анализа выделяются соответствующие онтологические знания, обеспечивающие эффективное решение, и выстраиваются механизмы взаимодействия знаний. Ниже рассматриваются некоторые важные задачи лингвистического анализа и подходы к их решению.

a. Разрешение грамматической и лексической многозначности. Задача разрешения многозначности является центральной задачей лингвистического анализа. Многозначность является неотъемлемым свойством естественного языка и присуща языковым явлениям всех уровней. Разрешение грамматической (функциональной) омонимии является этапом синтаксического анализа, разрешение лексической омонимии используется на этапе построения семантических интерпретаций текста. Традиционно, методы разрешения функциональной омонимии разрабатываются в рамках статистического подхода или подхода, основанного на правилах. Для эффективного разрешения омонимии также можно использовать ряд специальных ресурсов (онтологии, базы фильтров и др.). Эти ресурсы поддерживаются онтологической компонентой онтолого-лингвистической системы и доступны на различных этапах лингвистического анализа текста.

b. Сегментация предложения (синтаксический анализ). Использование результатов онтологической разметки в методах сег-

Рис. 1. Типовые функциональные компоненты онтолого-лингвистической системы

ментации предложения позволяет эффективно выделять границы сегментов. Специализированные ресурсы онтологической компоненты (специальные базы сегментов) также улучшают результаты сегментации.

с. Разрешение референции и восстановление эллипсиса. Механизмы референции и восстановления эллипсиса некоторых типов могут эффективно базироваться на онтологиях и специальных онтологических выводах.

4. Поддержка онтологических выводов. Онтологическая компонента осуществляет поддержку различных типов выводов по онтологии для построения решения различных задач.

Лингвистическая компонента обеспечивает решение следующих лингвистических задач обработки текстов:

– распознавание символов (токенизация). Традиционная задача обработки текстов, связанная с распознаванием и классификацией объектов текста (словоформы, числа, знаки препинания и др.);

– сегментация предложений. Выделение границ предложений в тексте решается на основе правил распознавания функций знаков препинания; (внешние или внутренние знаки).

– распознавание типов лингвистических объектов (словоформа, число, дата, время, аббревиатура и т.п.). На данном этапе уточняется подтип выделенного на этапе токенизации лингвистического объекта. Распознавание осуществляется на основе правил и специальных словарных ресурсов;

– морфологический анализ словоформ. Классическая задача обработки текстов решается на основе грамматических словарей и бессловарными методами с генерацией множества гипотез о лексико-грамматических разрядах и способах словоизменения известных и неизвестных слов на основе правил;

– разрешение грамматической и лексической многозначности. Центральная задача анализа текста решается на основе статистических или контекстных методов. Для разрешения лексической многозначности используются словари сочетаемости, а также онтологические методы;

– синтаксический анализ и разрешение синтаксической многозначности. Сложность задачи синтаксического анализа связана с синтаксической многозначностью и комбинаторным взрывом при переборных решениях. Для сокращения перебора используются различные механизмы фильтрации. Подходы к синтаксическому анализу различаются формальными моделями синтаксиса;

– разрешение референции и восстановление эллипсиса. Задача особенно актуальна для приложений в области извлечения семантической информации из текстов.

Основные задачи построения онтолого-лингвистической системы для приложений компьютерного анализа текстов можно условно разделить на задачи лингвистического анализа текстов, задачи проектирования онтологий, задачи подготовки ресурсов и построения решений прикладных задач.

Сложность задач лингвистического анализа текстов связана с неодинаковой разработанностью алгоритмического обеспечения различных этапов лингвистического анализа. Разработка алгоритмического обеспечения опирается на формальные модели, при этом разные этапы лингвистического анализа текстов формализованы в разной степени. Наиболее формализованным является морфологический уровень, который описывается различными формализмами (модели конечных автоматов, регулярные грамматики и др.). Однако и для этого уровня можно указать сложные задачи

морфологического и особенно морфосинтаксического анализа, требующие дальнейших исследований. Алгоритмы синтаксического анализа базируются на различных синтаксических теориях [7-12]. Однако синтаксические теории, как правило, описывают макроуровень синтаксиса, в то время как пограничные явления (морфосинтаксические, семантико-синтаксические, микросинтаксис) остаются недостаточно исследованными [23].

Несомненно, более сложной является задача моделирования семантики. Язык является сложнейшим объектом для семантического моделирования. Назовем лишь некоторые из наиболее фундаментальных свойств естественного языка: принципиальная нечеткость значения языковых выражений; динамичность языковой системы; образность номинаций, основанная, прежде всего, на метафоричности; креативность в освоении новых знаний; семантическая мощь словаря, позволяющая выражать любую информацию с помощью конечного множества элементов; гибкость в передаче информации; разнообразие функций; специфическая системность.

Можно выделить ряд направлений, в которых развиваются формальные семантические теории: лексическая семантика, формальная семантика, вычислительная семантика, когнитивная семантика, онтологическая семантика.

Онтологическая семантика – относительно новое направление формальной семантики, связанное с разработкой формальных моделей представления онтологических знаний и построения онтологических выводов, интерпретация языка в терминах языковонезависимых концептов проблемной области.

Проблемы онтологической семантики обсуждаются здесь в связи с фундаментальной проблемой построения онтологий и использованием онтологических знаний в лингвистических приложениях. К наиболее адекватным онтологическим моделям, применимым в задачах лингвистического анализа, в первую очередь относятся лексические (лингвистические) онтологии. Лексические онтологии строятся как иерархические лексические ресурсы (типа WordNet). Единицей в таких системах является значение отдельной лексемы или синсета (совокупности синонимов). Между единицами устанавливаются различные типы отношений, прежде всего иерархические отношения. Лексические онтологии пред-

ставляют собой достаточно узкий класс онтологических ресурсов и общей проблемой, широко обсуждаемой в литературе, является сопряжение лексических и формально-логических онтологических моделей в прикладных задачах.

Можно отметить ряд основных подходов к решению этой задачи:

– первичность разработки логической онтологической модели и затем сопоставление онтологическим концептам языковых значений [24, 25];

– первичность иерархического лексического ресурса (например, типа WordNet) и затем сопоставление лексических единиц онтологическим концептам некоторой онтологии [26];

– сбалансированное описание онтологических понятий и лексических значений [18].

К важнейшей задаче разработки онтолого-лингвистической системы относится подготовка информационных ресурсов (общих и специальных), обеспечивающих функционирование системы. Многие из существующих словарей требуют существенной доработки для встраивания их в компьютерные технологии. Разработка новых лингвистических ресурсов является крайне трудоемким и ресурсно затратным процессом. Существенное улучшение данной ситуации связывается с разработкой больших корпусов национальных языков, дающих богатый материал для построения различных лексических словарей и баз данных.

Построение решений прикладных задач выполняется под управлением системы онтологических моделей, что является отличительной особенностью онтолого-лингвистической системы.

2.3. СИСТЕМА ОНТОЛОГИЧЕСКИХ МОДЕЛЕЙ КАК ЯДРО ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ

Концепция онтолого-лингвистической системы реализована в системе «OntoIntegrator» [27]. Процесс решения прикладной лингвистической задачи в онтолого-лингвистической системе реализуется под управлением системы онтологических моделей [28]. Система онтологических моделей включает различные типы онтологий: прикладные онтологии, онтологию моделей и онтологию планирования задач. Структура системы онтологических моделей представлена на рисунке 2.

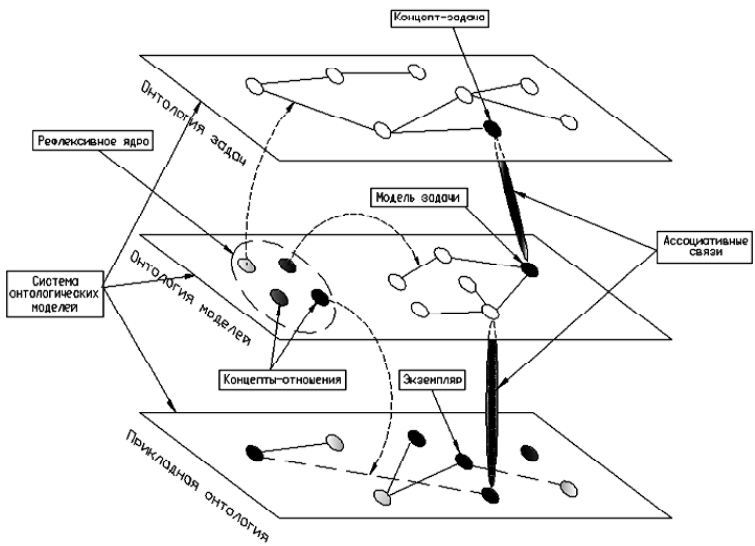

Рис. 2. Структура системы онтологических моделей

С точки зрения структурной организации система онтологических моделей представляет собой трехкомпонентную ассоциативную систему. Компонентами системы являются семантические сети (онтологические подсистемы): онтология планирования задач, онтология моделей и прикладные онтологии. Формально, трехкомпонентная онтологическая система есть структура вида $S = (\Psi, \Sigma, \Omega.)$ где Ψ – онтология планирования задач, Σ – онтология моделей, Ω – прикладная онтология.

Онтология проектирования задач $\Psi = (\Psi_K, \Psi_R, Z_\Psi)$ есть семантическая сеть, в которой Ψ_K множество концептов; Ψ_R – множество отношений, в общем случае n-местных $\Psi_R \subseteq \Psi^n_K$; Z_Ψ – множество функций интерпретации.

Онтология моделей $\Sigma = (\Sigma_K, \Sigma_R, H_R, Z_\Sigma)$ есть семантическая сеть, в которой Σ_K – множество концептов; Σ_R – множество отношений, в общем случае n-местных $\Sigma_R \subseteq \Sigma^n_K$; Z_Σ – множество функций интерпретации, H_R – рефлексивное ядро (подробно будет описано в п. 2.4).

Между компонентами (семантическими сетями) онтологической системы могут быть установлены ассоциативные связи (ас-

социативные отношения) R, $R \subseteq \Psi \times \Sigma$, $R \subseteq \Sigma \times \Omega$. Ассоциативные отношения устанавливаются между подсистемами соответствующих онтологий.

2.4. ОНТОЛОГИЯ ПЛАНИРОВАНИЯ ЗАДАЧ

Онтология планирования задач формально определяется как структура вида $\Psi = (\Psi_K, \Psi_R, Z_\Psi)$, где Ψ_K – множеств концептов-задач, Ψ_R – множество отношений, Z_Ψ – множество функций интерпретации. Концепт-задача $\psi\ (A_1, A_2, A_3) \in \Psi_K$ описывается набором атрибутов A_1, A_2, A_3. Атрибут A_1 определяет имя задачи, атрибут A_2 – функциональный тип задачи, атрибут A_3 – дескриптор задачи.

Правила объявления атрибутов задают имена атрибутов, функциональные типы и дескрипторы, а также устанавливаемые по умолчанию значения атрибутов. Имена атрибутов являются символьными данными. Дескриптор задачи вводит текстовое описание задачи (краткая аннотация задачи). Функциональный тип присваивает значение *функционального* класса задачи из открытого множества допустимых функциональных классов: *Операции (TOperations), Источники (TSources), Приемники (TSinks), ψ-Реализации (Tψ -Realizations)*.

Ниже приведен пример построения экземпляра класса *ψ-Реализации* в виде последовательности концептов задач.

Концепт-задача *K «Оценка функциональной омонимии текста»* включает линейную последовательность концептов-задач *K1, K2,.., K9*:

Концепт-задача K «Оценка функциональной омонимии текста»:

K1: TSinks // загрузка исходного текста

K2: TSinks //просмотр исходного текста

K3: TSinks // загрузка моделей функциональной омонимии

K4: TSources // лексический анализ текста

K5: TSinks //просмотр построенных моделей текста

K6: TSources // получение статистики по типам омонимии

K7: TSources // получение статистики по частотности омонимов

K8: TSources // получение статистики по частотности словоформ текста

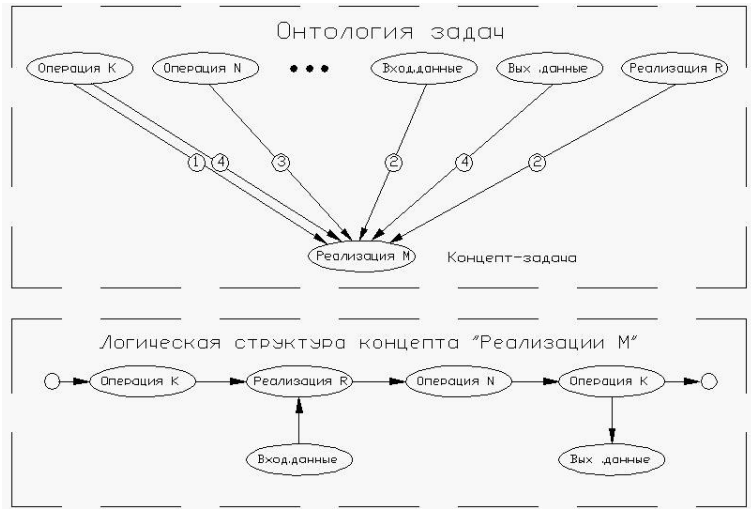

Рис. 3. Конструирование концепта-задачи

K9: TSinks // просмотр построенных моделей текста

Класс ψ-*Реализации* определяет структуру решения задачи в виде последовательности базовых операций и ψ-реализаций. Экземпляр класса ψ-*Реализации* представляет собой исполняемый модуль для решения определенной прикладной задачи. Класс *Операции* содержит расширяемый набор базовых операций, для которых фиксируется конкретная семантика. Классы *Источники* и *Приемники* определяют средства преобразования данных, различающиеся функциональностью. Источники обеспечивают доступ к содержимому и его преобразование. Приемники обеспечивают загрузку содержимого в память (во внутренние структуры данных).

Формирование экземпляра класса ψ-*Реализации* происходит на основе специального механизма назначения последовательности операций (планирование логической структуры реализации) на основе отношения включения из множества Ψ_R отношений онтологии планирования задач. Отношение включения реализуется с атрибутом «номер следования», метрика отношения используется для передачи параметров между операциями.

На рисунке 3 представлен процесс назначения последовательности применяемых операций и реализаций при конструи-

ровании новой ψ-*Реализации* и логическая структура созданной ψ-*Реализации.* Конструирование нового концепта-задачи происходит на основе выбранных *операций, реализаций и источников данных*. Формирование логической структуры концепта «*Реализация N»* базируется на механизме приписывания выбранным *операциям* и *реализациям* номеров следования в общей логической последовательности. Так, в рассматриваемом примере «*операции K*» присвоены номера 1 и 4, «*операции N*» – номер 3, «*Реализации R*» и «*Источники*» – номера 2, «*Приемники*» – номер 4.

2.5. ПРИКЛАДНЫЕ ЛИНГВИСТИЧЕСКИЕ ОНТОЛОГИИ В ОНТОЛОГО-ЛИНГВИСТИЧЕСКОЙ СИСТЕМЕ

Согласно современным представлениям, термину «онтология» удовлетворяет широкий спектр структур, представляющих знания о той или иной предметной области. В качестве в разной степени формализованных онтологий рассматриваются [29]:

1) словарь с определениями;
2) простая таксономия;
3) тезаурус (таксономия с терминами);
4) модель с произвольным набором отношений;
5) таксономия и произвольный набор отношений;
6) полностью аксиоматизированная теория.

При этом различаются [30] фундаментальные онтологии (*fundamental ontologies*), которые описывают предметную область максимально полно (пункт 6 в списке выше), безотносительно к приложениям и обычно с максимальной степенью формализации, и прикладные онтологии (*application ontologies*), которые также называются «легкими» онтологиями (*lightweight ontologies*) и которые формализуются настолько, насколько это необходимо для приложения (пп. 1-5 в списке выше). Общим для всех формализаций является выделение множества объектов (концептов, понятий), алфавита отношений, правил установления отношений и аксиом, задающих правила вывода на множестве отношений.

Для приложений, ориентированных на различные задачи текстового анализа, в том числе, информационный поиск, особое значение имеют лингвистические онтологии. Главной характеристикой лингвистических онтологий является то, что они связа-

ны со значениями языковых выражений (слов, именных групп и т.п.). При построении предметных онтологий, как правило, сначала строится система понятий, которым затем приписываются наборы языковых выражений (слов, терминов, словосочетаний). Особенность лингвистических онтологий проявляется прежде всего в том, что понятийный аппарат онтологии базируется на лексической системе языка и лексических значениях. Лексическое значение лежит в основе разделения понятий и группа понятий с общим лексическим значением образует группу квазисинонимов, называемых текстовыми эквивалентами (текстовыми входами) понятия.

Можно отметить ряд крупных проектов, связанных с разработкой лингвистических онтологий:

– ресурс WordNet [31], представляющий в виде иерархической структуры систему значений слов общезначимого английского языка. На основе модели WordNet традиционно создаются терминологические системы конкретных предметных областей, то есть лингвистические онтологии этих областей [32, 33];

– онтология Mikrokosmos [34], в которой предложен сбалансированный подход к описанию системы значений языковых единиц и связанной с ними системы понятий;

– РуТез*Онтологии, включая Тезаурус русского языка РуТез и онтологию по естественным наукам и технологиям для приложений в сфере информационного поиска [35, 36]. Оба ресурса принадлежат к классу лингвистических онтологий, но, в то же время, являются и тезаурусами, поскольку каждое понятие связано с набором языковых выражений (слов, терминов, словосочетаний), которыми это понятие может быть выражено в тексте.

Формально, онтология может быть представлена как структура вида [29] $O = < X, R, F >$, где X – конечное множество концептов (понятий, терминов) предметной области онтологии; R – конечное множество отношений между концептами (понятиями, терминами) заданной предметной области; F – конечное множество функций интерпретации (аксиоматизации).

Лингвистическую онтологию O^L можно формально определить как структуру вида $O^L = < X, S^X, R, F >$, где X – конечное множество концептов предметной области онтологии; S^X конечное

множество текстовых эквивалентов (квазисинонимов) понятий из множества X; $R = < R_X, R_S >$ – конечное множество отношений; F – конечное множество функций интерпретации (аксиоматизации).

Множество S^X содержит объекты (лексемы, словосочетания), которые условно интерпретируются как текстовые эквиваленты (квазисинонимы) онтологических понятий. Фактически, распознавание в тексте онтологического понятия реализуется на основе механизма распознавания текстовых эквивалентов.

Множество отношений R состоит из двух множеств: $R_X \subseteq X \times X$ – множество отношений между концептами предметной области; $R_S \subseteq X \times S^X$ – множество отношений между концептами и квазисинонимами. Множество отношений R_X задается при проектировании онтологии и выбор конкретных отношений, связывающих концепты предметной области, определяется целью и задачами проектирования.

$R_S = \bigcup_{i \in X} R_S^i$, где $R_S^i = \{< x^i, s^k >, k \in K_i\}$ – множество квазисинонимов онтологического понятия x^i. Отношение R_S^i устанавливает ассоциативную связь между понятием x^i и набором текстовых эквивалентов (квазисинонимов) понятия s^1, s^1, ... , s^k.

Разработанные совместно с Добровым Б. В. и Лукашевич Н. В. методы проектирования лингвистических онтологий изложены в [37].

Проектирование лингвистических онтологий включает несколько этапов:

1) формирование текстовой коллекции заданной предметной области. Коллекция должна быть представительной и включать учебники, научные статьи, техническую документацию, материалы средств массовой информации и др.;

2) построение списка терминов предметной области на основе методов извлечения терминов из текстовой коллекции.

В настоящее время можно выделить три основных класса методов извлечения терминологии из текстов: лингвистические методы, статистические методы и комбинированные методы [38].

Лингвистические методы базируются на лексико-синтаксических шаблонах однословных и многословных терминов и системе фильтров для отсеивания «нетерминов».

Применение статистических методов опирается на представление о частотности терминов. Терминосочетания обычно соотносятся с *n*-граммами (двух-, трех-, четырехчленными сочетаниями), характеризующимися высокой степенью устойчивости. В качестве мер, пригодных для оценки устойчивости словосочетаний в текстах, обычно используются *MI-score*, *t-score*, *Log-Likelihood*, *C-value*, критерий χ^2 и ряд других.

Комбинированные методы анализа терминологии предполагают совместное использование аппарата лексико-грамматических шаблонов, методов сборки терминосочетаний, системы фильтров, а так же статистического аппарата.

Построение сети онтологических понятий выполняется на основе выбранной системы онтологических отношений. Онтологические понятия выделяются на основе построенного списка терминов предметной области. Выделение понятия основывается на тщательной экспертизе, которая учитывает ряд важных факторов, таких как, значимость, лексическая многозначность, отнесенность термина к разрядам общей или специальной лексики, анализ состава словосочетания для многословных понятий и др.

Для каждого понятия формируется список текстовых эквивалентов (квазисинонимов). Затем осуществляется построение сети онтологических понятий с помощью системы выбранных отношений. Набор отношений в большой степени зависит от предметной области и от задачи, для решения которой предназначена онтология. Однако, можно выделить набор универсальных отношений, не зависящих от предметной области и типа решаемой задачи. В онтологической семантике выделяются универсальные отношения иерархии (класс-подкласс), отношения части-целого, отношения принадлежности классу (элемент-класс), отношения онтологической зависимости [39]. Если существование понятия X зависит от существования понятия Y, то будем говорить, что понятие X находится в отношении онтологической зависимости от понятия Y. Конкретная семантика отношения онтологической зависимости может быть различной, например, различают отношения строгой зависимости (*rigid dependence*), родовой зависимости (*generic dependence*), исторической зависимости и др.

Отношения онтологической зависимости могут быть семантически специфицированы для конкретных предметных областей и приложений.

Приведем в качестве примера статью понятия *пуск ракеты* из лингвистической онтологии авиационной области [40], разработанной совместно с Лукашевич Н. В.

Концепт **ПУСК РАКЕТЫ**

син	*запуск ракеты // текстовый эквивалент*
син	*запустить ракету*
син	*отстрел ракеты*
син	*применение ракет*
син	*применение ракетного вооружения*
син	*пуск ракеты*
син	*пустить ракету*
син	*ракетная атака*
син	*ракетный удар*
ВЫШЕ	*ПРИМЕНЕНИЕ ОРУЖИЯ // отношение иерархии*
НИЖЕ	*ЭФФЕКТИВНЫЙ ПУСК РАКЕТЫ // отношение иерархии*
ЧАСТЬ	*ДАЛЬНОСТЬ ПУСКА РАКЕТЫ // отношение часть-целое*
ЧАСТЬ	*ЗОНА ВОЗМОЖНОГО ПУСКА*
ЧАСТЬ	*МАКСИМАЛЬНО ДОПУСТИМАЯ ПЕРЕГРУЗКА ПРИ ПУСКЕ*
АСЦ1	*РАКЕТА // отношение ассоциации «зависит_от» (отношение онтологической зависимости)*
АСЦ2	*ВЫХОД В ЗОНУ ВОЗМОЖНОГО ПУСКА // отношение ассоциации «главное_для» (отношение онтологической зависимости)*
АСЦ2	*КОМАНДА «ПУСК РАЗРЕШЕН»*
АСЦ2	*ОШИБКА ПУСКА*
АСЦ2	*УПРАВЛЕНИЕ РАКЕТОЙ*

2.6. МОДЕЛЬ ЛИНГВИСТИЧЕСКОЙ ОБОЛОЧКИ ОНТОЛОГИИ

В структуре онтолого-лингвистической системы выделяются две основные взаимодействующие компоненты: ***онтологическая*** и ***лингвистическая***. ***Онтологическая компонента*** поддерживает проектирование системы онтологических моделей и обеспечи-

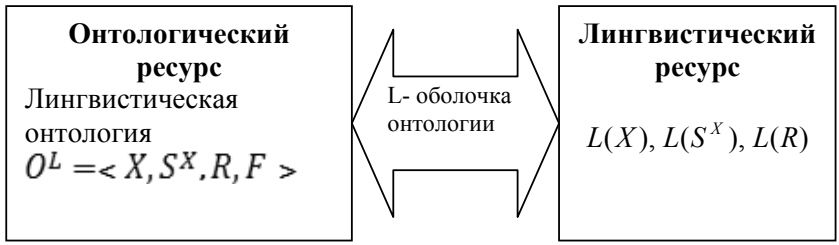

Рис. 4. Лингвистическая оболочка онтологии

вает взаимодействие с лингвистической компонентой при разработке лингвистических приложений. Одной из важных задач является сопряжение онтологических и лингвистических моделей представления знаний. Онтологическая модель задает способ структурирования знаний, лингвистическая модель определяет способы функционирования языковых единиц в текстах. Другими словами, лингвистическая модель определяет лексические и синтаксические модели концептуальных единиц, в том числе, онтологических единиц. Интерфейсом между онтологической и лингвистической моделями служит модель лингвистической оболочки онтологии [41].

На рисунке 4 обозначены *L(X)* – лингвистическая модель элементов множества *X, L(S^X)* – лингвистическая модель элементов множества S^X, *L(R)* – лингвистическая модель элементов множества *R*.

Введем формальное определение понятия лингвистической оболочки онтологии (*L*-оболочка онтологии). Для лингвистической онтологии *L*-оболочка онтологии задается структурой множеств вида $L = < X^L, G^L >$, где X^L – множество объектов *L*-оболочки, G^L – множество атрибутов объектов. Множество атрибутов есть декартово произведение множеств вида $G^L = G^{lex} \times G^{syn} \times G^{pos} \times G^{comp} \times G^d$, где G^{lex} – множество лексических параметров, G^{syn} – множество синтаксических параметров, G^{pos} – множество грамматических категорий, G^{comp} – множество композиционных параметров, G^d – множество дистантных параметров. Множество лексических параметров $G^{lex} = G_1^{lex} \times G_2^{lex} \times \cdots \times G_n^{lex}$, где G_i^{lex} – множество грамматических значений (граммем) грамматической категории G_i. Число грамматических категорий (n) определяется частереч-

ной принадлежностью лексемы, т.е. значением параметра из множества G^{pos}. Множество синтаксических параметров G^{syn} задают параметры синтаксического согласования слов в словосочетании. Множество композиционных параметров G^{comp} определяют параметры правил композиции многословного термина, включая параметры вариативности составляющих многословного термина. Множество дистантных параметров G^{d} задает параметры, определяющие расстояния между составляющими многословного термина.

Множество объектов *L*-оболочки X^L включает понятия онтологии, выраженные однословными или многословными терминами, а также списки текстовых эквивалентов (квазисинонимов) онтологических понятий. Таким образом, в *L*-оболочке с каждым онтологическим понятием $x \in X^L$ соотносится элементарный кортеж параметров $g_x = (g_x^{lex}, g_x^{syn}, g_x^{pos}, g_x^{comp}, g_x^{d})$. С каждым

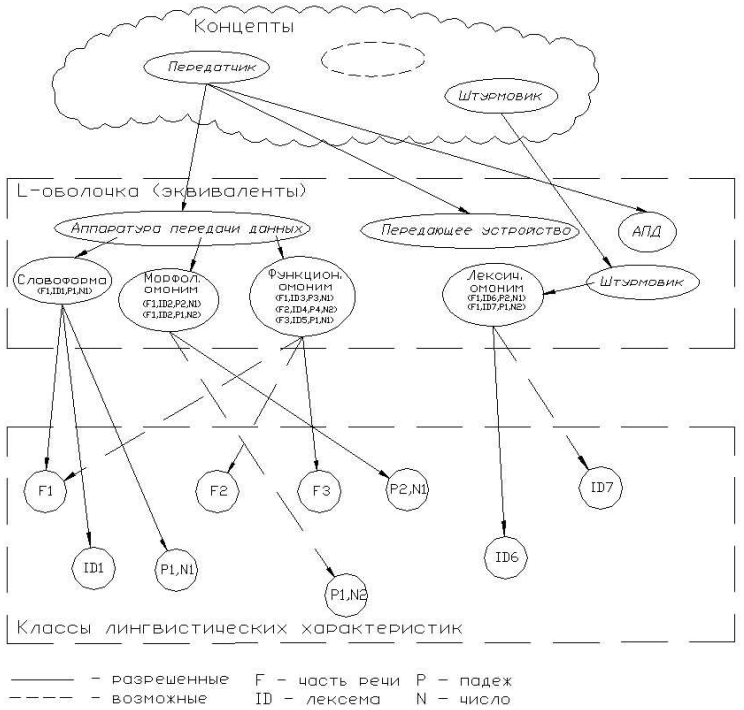

Рис. 5. Структура *L*-оболочки

многозначным понятием соотносится некоторый набор элементарных кортежей.

Формально, *L*-оболочка есть отображение множества объектов X^L на множество параметров G^L.

При этом одному объекту $x \in X^L$ могут соответствовать разные наборы параметров из G^L (частеречная омонимия), а также одному набору параметров из G^L может соответствовать несколько объектов из X^L (параметр класса объектов).

На рисунке 5 схематично представлен фрагмент *L*-оболочки с набором концептом и их описаний. В составе текстовых эквивалентов концептов имеются омонимы (функциональные и лексические), которые задаются соответствующим набором возможных грамматических параметров (среди которых выделен набор разрешенных параметров).

Метод встраивания онтологии в онтолого-лингвистическую систему на основе *L*-оболочки реализован в специальном программном модуле системы «OntoIntegrator». Таким образом, *L*-оболочка онтологии обеспечивает технологическую основу для проектирования методов распознавания онтологических понятий в текстах и разрешения многозначности, которые являются базовыми в задачах автоматической обработки текстов.

Заключение

В работе рассмотрен подход к проектированию онтолого-лингвистических систем, а также некоторые аспекты реализации предложенного подхода в онтолого-лингвистической системе «OntoIntegrator».

Класс онтолингвистических систем отличается объединением экстралингвистических (онтологических) и лингвистических знаний, эвристических и формальных методов обработки ЕЯ. Ядром онтолингвистических систем являются знания различной природы, в том числе, различные онтологии, представляющие прикладные знания, метазнания, в том числе, знания о прикладных задачах и их свойствах.

Основные технологические задачи онтолингвистических систем решаются через взаимодействия онтологической и лингвистической компонент системы. Соответствующие технологии взаимодействия позволяют решать общие и специальные задачи ана-

лиза текстов, при этом общая структура решаемой задачи может динамически меняться через специальные механизмы настройки типа решаемой задачи.

Центральным ядром онтолого-лингвистических систем является система онтологических моделей, под управлением которой выполняется построение решения прикладной задачи. Система онтологических моделей включает различные типы онтологий: прикладные онтологии, онтологию моделей и онтологию планирования задач.

В работе рассмотрены формальные модели описания структуры онтологии моделей и онтологии планирования задач. Введены типы концептов и отношений в соответствующих онтологиях. Разработаны методы формирования представления прикладных задач в соответствующих формализмах.

Разработанная формальная модель лингвистической оболочки онтологии, позволяет эффективно включать внешние прикладные онтологии в состав онтолого-лингвистической системы.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Kaiser K., Miksch S. Information Extraction. A Survey // Vienna University of Technology. Institute of Software Technology & Interactive Systems. Asgaard-TR-2005-6. [Электрон. ресурс]. – 2005. – URL: http://ieg.ifs.tuwien.ac.at/techreports/Asgaard-TR-2005-6.pdf (дата обращения: 10.09.2016).

2. Chia-Hui Chang, Mohammed Kayed M., Girgis M. R., Shaalan K. A Survey of Web Information Extraction Systems // *IEEE Transactions on knowledge and data engineering, TKDE-0475-1104.R3*. [Электрон. ресурс]. – 2006. – URL: http://in1.csie.ncu.edu.tw/~chia/pub/iesurvey2006.pdf (дата обращения: 10.09.2016).

3. Indurkhya N., Damerau F. J. Handbook of Natural Language Processing. – Chapman & Hall/CRC, 2010. – 692 pp.

4. Bunt H., Merlo P., Nivre J. (eds.) Trends in Parsing Technology. Dependency Parsing, Domain Adaptation, and Deep Parsing. – Springer, 2010. – 307 pp.

5. Hermann H.. Knowledge Representation and the Semantics of Natural Language. – Springer, 2006. – 651 pp.

6. Kao A., Roteet S. Natural Language Processing and Text Mining. – Springer, 2006. – 265 pp.

7. Matthews P. H. Syntactic Relations: a critical survey. – London: Cambridge University Press, 2007.

8. Mel′čuk I. A. Dependency syntax: theory and practice. – New York: Albany State University, 1987.

9. Mel′čuk, I. A. Levels of dependency in linguistic description: Concepts and problems // *Dependency and valency: An International handbook of contemporary research*, ed. Agel et al. – Berlin: Walter de Gruyter, 2003. – 170–187 pp.

10. Boguslavsky I., Iomdin L., Sizov V. Multilinguality in ETAP-3. Reuse of Linguistic Resources // *Proceedings of the Workshop "Multilingual Linguistic Resources". 20th International Conference on Computational Linguistics*. – Geneva: 2004. – P. 7–14.

11. Apresian J., Boguslavsky I., Iomdin L., Lazursky A., Sannikov V., Sizov V., Tsinman L. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT // *MTT 2003. First International Conference on Meaning-Text Theory*. – Paris: Ecole Normale Superieure, 2003. – P. 279–288.

12. Богуславский И. М., Диконов В. Г., Иомдин Л. Л., Тимошенко С. П. Как моделировать понимание естественного языка: формальное представление смысла // *Компьютерная лингвистика и интеллектуальные технологии*. По материалам ежегодной Международной конференции «Диалог» (2013). – Выпуск 12. – Том 2. – М.: Издательский центр «Российский государственный гуманитарный университет», 2013. – С. 132–144.

13. Java A., Nirenburg S., McShane M., Finin T., English J., Joshi A. Using a Natural Language Understanding System to Generate Semantic Web Content // *International Journal on Semantic Web and Information Systems*. – 2007. – Vol. 3(4). – P. 50–74.

14. Ge R., Raymond J. Mooney R. J. A Statistical Semantic Parser that Integrates Syntax and Semantics // *Proceedings of the Ninth Conference on Computational Natural Language Learning*. – Ann Arbor: MI, 2005. – P. 9–16.

15. Liang P., Jordan M., Klein D. Learning Dependency-Based Compositional Semantics // *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. – Vol. 1. – 2011. – P. 590–599.

16. Hitzler P., Krφtzsch M., Rudolph S. Foundations of Semantic Web Technologies. – CRC Press/Chapman and Hall, 2009.

17. Szeredi P., Lukαcsy G., Benkõ T. The Semantic Web Explained – the technology and mathematics behind Web 3.0. – London: Cambridge University Press, 2014.

18. Nirenburg S., and Raskin V. Ontological Semantics. – London: The MIT Press, Cambridge Mass, 2004.

19. Poon H., Domingos P. Unsupervised semantic parsing // *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. – Vol. 1. – EMNLP 09. – 2009. – P. 1–10.

20. Biemann C. Ontology Learning from Text: A Survey of Methods // *LDV forum*. – Vol. 20. – No. 2. – 2005. – P. 75–93.

21. Starostin A. S., Smurov I. M., Stepanova M. E. A Production System for Information Extraction Based on Complete Syntactic-Semantic Analysis // *Компьютерная лингвистика и интеллектуальные технологии*. По материалам ежегодной Международной конференции «Диалог-2014». – Вып. 13. – М.: Издательский центр «Российский государственный гуманитарный университет», 2014. – С. 659–667.

22. Borschev V. B., Partee B. H. Ontology and Integration of Formal and Lexical Semantics // *Компьютерная лингвистика и интеллектуальные технологии*. По материалам ежегодной Международной конференции «Диалог-2014». – Вып. 13. – М.: Издательский центр «Российский государственный гуманитарный университет», 2014. – С.114–127.

23. Иомдин Л. Л. В глубинах микросинтаксиса: один лексический класс синтаксических фразем // *Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2008*. – Вып. 7 (14). – М.: РГГУ, 2008. – С. 178-184.

24. Gruber T. R. A translation approach to portable ontologies // *Knowledge Acquisition*. – 1993. – Vol. 5(2). – P. 199–220.

25. Reed S., Lenat D. Mapping ontologies into CyC // *Proceedings of the AAAI'02 workshop on Ontologies and the Semantic Web*. – Edmonton, Canada: 2002.

26. Atserias J., Clament S., Rigau G. Toward the Meaning Top Ontology: Sources of Ontological Meaning // *Proceedings of International conference Language Resources and Evaluation (LREC-2004)*. – Vol. 1. – 2004. – P. 11–14.

27. Невзорова О. А., Невзоров В. Н. Интеллектуальная инструментальная система «OntoIntegrator» для задач автоматической обработки текстов // *Тринадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2012 (16–20 октября 2012 г., г. Белгород, Россия): Труды конференции.* – Т. 4. – Белгород: Изд-во БГТУ, 2012. – С. 92–99.

28. Невзорова О. А., Невзоров В. Н. Многоуровневая онтологическая система для планирования решений прикладных задач // *Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2011): материалы Междунар. научн.-техн. конф, Минск, 10–12 февраля 2011*. – Минск: БГУИР, 2011. – С. 323–330.

29. Гаврилова Т. А., Хорошевский В. Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2000.

30. Gomez-Perez A., Fernandez-Lopez M., Corcho O. OntoWeb // *Technical Roadmap. D.1.1.2. – IST project IST-2000-29243*. – [Электрон. ресурс]. – 2000. – URL: www.aifb.uni-karlsruhe.de/WBS/ysu/publications/OntoWeb_Del_1-1-2.pdf (дата обращения: 10.09.2016).

31. Miller G., Beckwith R., Fellbaum C., Gross D., Miller K. Five papers on WordNet // *CSL Report 43*. – Cognitive Science Laboratory, Princeton University, 1990.

32. Buitellar P., Sacalenau B. Extending Synsets ith Medical Terms // *Proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. – Pittsburg, USA: 2001.

33. Vossen P., Extending, Trimming and Fusing WordNet for Technical Documents // *Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. – Pittsburg, USA: 2001.

34. Mahesh K., Nirenburg S. A Situated Ontology for Practical NLP // *Proc. Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95)*. – Montreal, Canada: 1995.

35. Добров Б. В., Лукашевич Н. В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска // *Ученые записки Казанского гос. ун-та. Сер. «Физ.-матем. науки»*. – Т. 49. – № 2. – Казань: Изд-во Казанского ун-та, 2007. – С. 49–72.

36. Лукашевич Н. В., Добров Б. В., Тезаурус русского языка для автоматической обработки больших текстовых коллекций // *Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002* / Под ред. А. С. Нариньяни. – Т. 2. – М.: Наука: 2002. – С. 338–346.

37. Добров Б. В., Лукашевич Н. В., Невзорова О. А., Федунов Б. Е. Методы и средства автоматизированного проектирования прикладной онтологии // *Известия Российской академии наук. Теория и системы управления*. – № 2. – 2004. – С. 58–68.

38. Лукашевич Н. В., Логачев А. М. Комбинирование признаков для автоматического извлечения терминов // *Вычислительные методы и программирование*. – Т. 11. – 2010. – С. 108–116. – [Электрон. ресурс]. – 2010. – URL: http://num-meth.srcc.msu.su/ (дата обращения: 10.09.2016).

39. Guarino N. Some ontological principles for designing upper level lexical resources // *Language Resources and Evaluation*, A. Rubio,

N. Gallardo, R. Castro, A. Tejada (Eds.), European Language Resources Association (ELRA). – Vol. 1. – Granada, Spain: 1998. – P. 527–534.

40. Лукашевич Н. В., Невзорова О. А. АвиаОнтология: анализ современного состояния ресурса // *Труды международной конференции Диалог 2004 «Компьютерная лингвистика и интеллектуальные технологии»*. – М.: Наука, 2004. – С. 424–430.

41. Невзорова О. А. Онтолингвистические системы: технологии взаимодействия с прикладной онтологией // *Ученые записки Казанского государственного университета. Серия «Физико-математические науки»*. – Том 149. – Кн. 2. – 2007. – С. 105–115.

Глава 3. СОЗДАНИЕ ПЕРСОНАЛЬНОЙ ОНТОЛОГИИ

П. И. Соснин

3.1. ВВЕДЕНИЕ

Деятельность – это форма жизни, в основу которой природа положила интеллектуально обработанные условные рефлексы, формирующие то, что позволяет ввести в типовые взаимодействия человека с окружающей средой «вторую сигнальную систему» [1], то есть естественный язык (ЕЯ). Приписывание знаков ЕЯ обстоятельствам, которые распознаются с помощью органов чувств в определённой окружающей обстановке, открывает возможность для абстрагирования от этой обстановки до знакового образца, позволяющего управлять соответствующим типовым поведением человека.

Такое типовое поведение принято называть «прецедентом», а знаковый образец, по которому оно осуществляется «моделью прецедента». То, что называют «опытом человека» включает громадное количество «прецедентов», взаимодействие с которыми (для развития, поиска, доступа, настройки, рационализации и других целей) невозможно без их систематизации.

Разумно различать неявную естественную систематизацию, за существование которой отвечают естественные механизмы опыта человека, естественно-искусственную систематизацию, ответственность за которую несёт освоенная человеком система понятий, и искусственную систематизацию, материализация и использование которой осуществляется с помощью искусственных сред, например, компьютеризованного типа.

Первые два вида систематизации имеют отношение только к человеку, а третий вид допускает его применение не только человеком, а, например, и программными (интеллектуальными) агентами, в различных приложениях. Именно к третьему виду

относятся искусственные средства систематизации, получившие название «онтологий». За созданием любой «онтологии» всегда стоят задачи, для решения которых используются взаимодействия с опытом и его моделями.

В работе предлагается подход к формированию персональных онтологий, обслуживающих взаимодействие конкретного индивида с доступным ему опытом. Подход материализован инструментально в версии OwnWIQA вопросно-ответной моделирующей среды WIQA (Working In Questions and Answers) [2].

3.2. ПРЕДВАРИТЕЛЬНЫЕ ОСНОВЫ

Предлагаемый подход развивает наши исследования в проблемной области «Формирование и использование прикладных онтологий», результаты которых раскрыты в публикациях [3-4].

Специфику этих исследований, ориентированных на прикладные онтологии в проектировании автоматизированных систем (АС), определяют следующие установки:

1. Онтология конкретной прикладной области должна строиться в процессах проектирования АС, относящихся к этой области.

2. Функции основного информационного источника в построениях прикладных онтологий должны выполнять рассуждения лиц, вовлечённых в процессы проектирования и использования их результатов.

3. То, что включается в онтологию, должно прямо или опосредовано найти материальное воплощение в процессах проектирования и/или их продуктах.

4. Включение онтологии в состав средств, используемых в разработках АС и их применениях, должно конструктивно приносить ожидаемые позитивные эффекты.

Отметим, что среди различных типов рассуждений был выделен класс вопросно-ответных рассуждений (QA-рассуждений), осуществляемых в процессах решения проектных задач. Выбор этого класса был обусловлен диалоговой природой сознания, обеспечивающего доступ к наличному опыту.

Именно отмеченная специфика установок привела к созданию инструментально-технологического комплекса WIQA.Net, на базе которого разработан ряд его приложений, например, для концептуального проектирования АС, документирования и обучающего

сопровождения. В концептуальном проектировании «онтология проекта» применялась в оперативном предикатно-онтологическом контроле проектных решений [2].

Еще одним приложением, разработанном на основе WIQA.Net и использующем этот комплекс, является «База Опыта проектной организации», в состав которой включена и её онтология.

Так что реальные построения и применения прикладных онтологий привели к решению исследовать возможности построения и использования персональных онтологий. Представленные выше установки были сохранены, но в их приложении не только к задачам проектирования, которые по силам одному человеку, но и к освоению опыта решения профессиональных задач, с которыми сталкивается конкретный индивид. Этот класс задач в разработке персонального инструментария OwnWIQA было решено считать подобными задачам проектирования, которые можно решать в операционной среде OwnWIQA.

К установкам, модифицированным для построения OwnWIQA, была добавлена ещё одна:

5. Владелец персональной онтологии является и разработчиком и пользователем продуктов, при создании которых он применяет построенную (и освоенную) им онтологию.

На разработку инструментария OwnWIQA, а также методологического обеспечения его применений существенное воздействие оказали публикации в области эмпирической программной инженерии, в первую очередь, раскрывающие моделирование и использование профессионального опыта [5], вопросно-ответных механизмов доступа к базам опыта [6] и научного подхода к экспериментированию в человеко-компьютерных средах [7].

3.3. ВОПРОСНО-ОТВЕТНАЯ ПАМЯТЬ

В соответствии с установками предлагаемого подхода персональная «онтология» предназначена для систематизации моделей прецедентов, которые освоены индивидом и используются им в решении задач $Z = \{Z_k\}$, относящихся к области его профессиональных интересов.

В спецификациях «онтологии» и её материализации принципиальное место занимает вопросно-ответная память (QA-память), обобщённо представленная на рисунке 1.

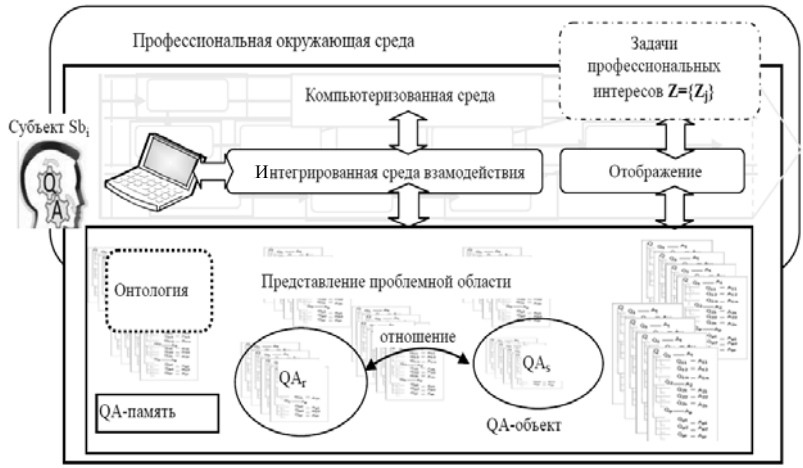

Рис. 1. Вопросно-ответная память инструментария OwnWIQA

QA-память – это подсистема инструментария OwnWIQA, предназначенная, в первую очередь, для моделирования QA-рассуждений, используемых индивидом (субъектом Sb_i) в процессах решения задач. Конкретная QA-модель состоит из связной совокупности моделей вопросов $\{Q_i\}$ и соответствующих им моделей ответов $\{A_i\}$, причём, каждая из таких моделей загружена в определённую ячейку QA-памяти.

Каждая из ячеек QA-памяти создаётся оперативно для регистрации и хранения атрибутики очередной модели типа Q или A. Атрибутика включает базовый набор атрибутов B(a1, a2, …, aN), к которому владелец OwnWIQA (или короче I) может добавить полезные дополнительные атрибуты AA(aa1, aa2, …, aaM). Совокупность базовых атрибутов с операциями над ними образует интерактивный объект типа Q или A. Ответственность за операции над объектом, в которых используется дополнительная атрибутика, возлагается на субъекта Sb_i.

В число базовых атрибутов ячейки QA-памяти включены: уникальное имя модели (тип и индекс, приписываемый автоматически), например Q1.1.2, способное выполнять роль адреса ячейки; идентификатор создателя модели, то есть имя I; знаковая модель Q или A, в форме символьной строки; момент времени создания

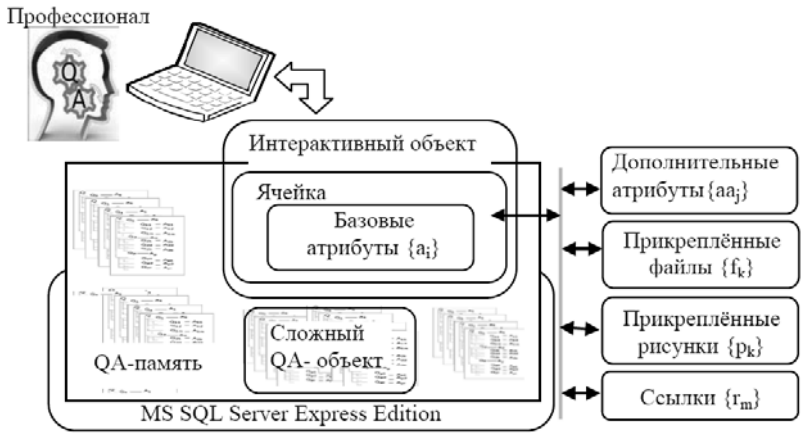

Рис. 2. Интерактивный Q-объект

или модификации; имя проекта и другие атрибуты. Обобщённое представление ячейки QA-памяти приведено на рисунке 2.

Приписывание дополнительных атрибутов и программирование дополнительных операций осуществляется субъектом Sb_i в процессах решения задач, если это полезно. Кроме того, в инструментарий OwnWIQA встроены средства для решения ряда системных задач, в решениях которых применены дополнительная атрибутика с дополнительными операциями. К числу таких задач относятся, например, задачи формирования документов по шаблонам на базе QA-рассуждений, зарегистрированных в QA-памяти.

Отметим, что к любому интерактивному объекту можно прикрепить файлы, например, графические файлы, которые будут включаться в документы.

И всё же особой целостностью обладает пара ячеек QA-памяти, в которых содержатся взаимно дополнительные Q-объект и A-объект. С такой парой в QA-памяти связывается QA-объект, позволяющий представлять «предикативность» и «актуальное членение» в знаковых описаниях, например, «обстоятельств окружающей среды». За такой целостностью «взаимодополнительность вопроса и ответа», что позволяет приписывать характеристику «да» (например, истинности или правдоподобия) согласованным текстовым представлениям вопроса Q_i и ответа A_i.

С другой стороны, любая коммуникативная единица, представленная знаками языка, например «предложение», открыта для её актуального членения на «тему» и «рему» [8], указывающую на то «информационно новое», для выражения которого предложение создано. Актуальное членение проще всего выразить и зарегистрировать с помощью согласованных вопроса Q_i и ответа A_i.

Следовательно, с помощью QA-объекта можно представить в QA-памяти «переменную» (Q-объект) и её «значение» (A-объект), как для любого простого типа, так и для любого составного типа. Более того, «переменной» определённого типа с помощью дополнительных атрибутов можно приписать «характеристики типа», общепринятые в языках программирования, а также полезные «семантические характеристики».

За приписыванием «переменной» её «значения» стоит «предложение», с которым (с помощью дополнительных атрибутов) можно связать определённую «модальность» или их совокупность, например, приписав «значению» его «неопределённость», например, с помощью вероятностной меры.

А значит, в QA-памяти информационно представимы «объекты» и «обстоятельства» окружающей среды, причём с привязкой ко времени и с учётом измерений и идентификации их характеристик. Отображение двух объектов среды и связи между ними в QA-памяти обобщённо представлены на рисунке 1.

Рис. 3. Составной QA-объект

В QA-памяти представимы не только объекты предметной области $Z = \{Z_k\}$, но и процессы, структурированные алгоритмически. Так, например, шаг алгоритма, выраженный в виде псевдокодового оператора, может быть загружен в поле описания Q-объекта. Ответом на такую интерпретацию вопроса логично считать результат выполнения оператора.

Отметим, что в общем случае «вопрос» может включать «подчиненные вопросы», что приводит к иерархически связанным совокупностям QA-объектов, представленных на рисунке 3.

Отметим, что структура и содержание QA-объектов, определяются теми (профессиональными) задачами, которые приходится решать субъекту Sb_i. В частности, в QA-памяти представимы «требования» (requirements) и их «спецификация» в «проектах» или «материализация» в «продуктах», в создание которых субъект Sb_i внёс или вносит свой вклад, применяя доступный ему опыт.

3.4. ОНТОЛОГИЗАЦИЯ ПЕРСОНАЛЬНОГО ОПЫТА

3.4.1. Ориентация на прецеденты

QA-память инструментальной среды OwnWIQA специфицирована и материализована так, чтобы в её ячейках можно было представлять решения задач $\{Z_k\}$, за которыми стоят «прецеденты», в осуществлении которых участвует субъект Sb_i. Более того, для «моделей прецедентов», которые хранятся в QA-памяти, должны использоваться нормативные формы, обобщённая структура которых приведена на рисунке 4.

Модель конкретного прецедента P_k включает: 1) текстовую составляющую P^T_k, в виде постановки задачи; 2) логическую составляющую P^L_k; 3) формула которой представлена на рисунке 4; 4) вопросно-ответную модель P^{QA}_k задачи Z_k; 5) графическое представление прецедента P^G_k; 6) исходный псевдокод P^I_k; 7) исполняемый код P^E_k.

Нормативная модель прецедента построена таким образом, чтобы она раскрывала концептуальное содержание задачи и представляла её концептуальное решение (псевдокод решения). По этой причине в модели принципиальное место занимают конструкции на естественно-профессиональном языке L^P предметной области $\{Z_k\}$. Основными из этих конструкций являются текст

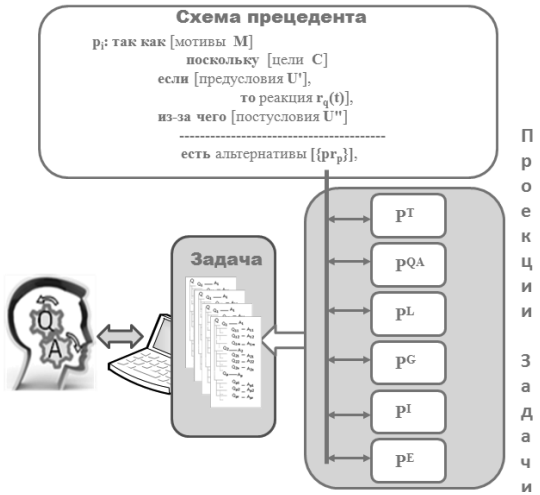

Рис. 4. Модель прецедента

T_k постановки задачи P^T_k и предложения $\{П_{ik}\}$ в модели P^{QA}_k, регистрирующей вопросно-ответный анализ этой задачи, в результате которого на все важные вопросы по её решению получены ответы, достаточные для построения логической схемы P^L_k и концептуально-алгоритмического решения P^I_k.

В комплексе OwnWIQA сохранена ориентация активности его пользователя на «проекты», что позволяет выделять на множестве задач $\{Z_k\}$ их взаимосвязанные совокупности $S_m\{Z_{mk}\}$, за каждой из которых стоит определённый «проект», выполненный или выполняемый субъектом Sb_i. Использование струкутуризации активности в единицах типа «проект» позволила заимствовать из комплекса WIQA.Net ряд средств для организации и управления процессами (псевдо) параллельного решения задач, а также заимствовать опыт создания и использования «онтологий проектов» [Соснин и др., 2012]. Так что в «персональной онтологии» $O_k = I(\{O_{mk}\})$ интегрируются совокупность «онтологий проектов» $\{O_{mk}\}$ субъекта Sb_i.

3.4.2. Лексика персональной онтологии

С множеством задач $\{Z_k\}$, решения которых ориентированы на прецеденты, связана определённая совокупность $\{T_k\}\cup\{\{П_{ik}\}\}$ употреблений языка LP. Именно этот лингвистический материал в

построениях персональной онтологии профессионального опыта должен занимать принципиальное место.

Следует отметить, что тексты и предложения из множеств $\{T_k\}\bigcup\{\{\Pi_{ik}\}\}$ создаются в результате согласованного уточнения их версий, в которых применялась не только лексика моделей прецедентов $\{P_k\}$, но и лексика естественного языка L в том объёме, которым владеет субъект Sb_i. А значит из языка L логично отбирать всё то, что будет полезным для решения задачи онтологической систематизации множества моделей $\{P_k\}$.

В комплексе QwnWIQA для такого отбора используется толковый «Словарь», который строится субъектом Sb_i в процессе создания системы моделей $S(\{P_k\}, O_k)$, включающей его персональную онтологию O_k. В этом словаре содержатся и понятия онтологии, требующих их определения. Так что лексика онтологии формируется из лексики моделей прецедентов и лексики, используемой в толковом «Словаре».

3.4.3. Формирование приращений онтологии

В структуре персональной онтологии O_k выделяются взаимосвязанные неформальная и формализованная составляющие. Неформальная часть состоит из понятий $\{N_q\}$ онтологии с их толкованиями (на базе языков L и L^P), представленными в статьях «Словаря».

В каждой статье второй части «имя статьи» N_q связывается со следующими списками:

- списком $(g_{q1}, g_{q2},... g_{qn})$ имён дифференцирующих признаков, согласованным с определением имени N;
- набором списков $\{(N_s, N_v,... N_w)\}$ имён других статей, каждый из которых связан с определённым типом систематизации моделей прецедентов.

Выбранная версия представления онтологии $O_k=I(\{O_{mk}\})$ упрощает её оперативное создание в процессе решения задач $\{S_m\{Z_{mk}\}\}$, так как типовым приращением для O_k является O_{mk}. Кроме того, каждая онтология O_{mk} развивается за счёт приращений, основными из которых являются имена объектов, зарегистрированных субъектом Sb_i в ячейках QA-памяти. По этой причине, формализованное содержимое онтологии O_{mk} адресуется, что упрощает его включение в программную обработку, например, для доступа к моделям прецедентов.

В инструментальной среде WIQA.Net для выявления приращений использовался предикатно-онтологический контроль текстовых описаний проектных решений в определённых «проектах». Предложения такого описания (автоматизированно) переводились на прологоподобный язык, после чего выделенные предикаты проходили проверку на соответствие текущему состоянию онтологии проекта. Отрицательный результат проверки указывал либо на ошибку, либо на необходимость ввести в онтологию приращение. Средства и механизмы такого контроля детально раскрыты в публикациях [2] и [4].

В разработке комплекса OwnWIQA было решено отказаться от выявления приращений онтологии за счёт предикатно-онтологического контроля в связи с его сложностью (использование визуализируемого псевдофизического силового взаимодействия между членами предложений в текстах описаний) для потенциальных пользователей OwnWIQA.

Однако визуализацию предикатной структуры предложений было решено сохранить, предоставив субъекту Sb_i возможности её графического (block-and-line) представления (с помощью специализированного графического редактора). Для выявления предикатной структуры субъект Sb_i может использовать морфологический анализатор, а в проверках механизмы онтологической фильтрации.

3.5. СРЕДСТВА СОЗДАНИЯ ПЕРСОНАЛЬНОЙ ОНТОЛОГИИ

Основным предназначением инструментария OwnWIQA является обслуживание концептуальной активности субъекта Sb_i в его профессиональной деятельности. Ориентация на профессиональную деятельность означает только то, что решаемые субъектом Sb_i задачи группируются и связываются в образования, которым можно приписать статус «проекта». Такая ориентация вводит связность в совокупности моделей прецедентов, а значит и системность в доступный для Sb_i опыт, так как часть этого опыта представлена моделями прецедентов.

Ориентация в OwnWIQA на разработку проекта привела к решению связать типовую единицу системности профессионального опыта с онтологией проекта, а интеграцию таких онтологий,

Рис. 5. Концептуальная структура онтологии

то есть интеграцию $\{O_{mk}\}$, связать с персональной онтологией O_k профессионального опыта субъекта Sb_i. Это же решение позволило отделить построение онтологий от действий по их интеграции, что и привело к нормативной концептуальной структуре «онтологии проекта», представленной на рисунке 5 в виде, который принято называть framework.

Для разделения онтологий по предметным областям, используются **словари**. Каждый словарь описывается следующими полями:

1. Название. Данное поле может использоваться как название предметной области, или, возможно, просто для названия набора понятий.

2. Описание. Данное поле может использоваться для более детального описания предметной области, которую описывает данный словарь.

Для описания разделов предметных областей используются группы понятий. Группы можно использовать также для простой классификации понятий, разбивая большое число понятий, на несколько не больших наборов. Группа понятий имеет следующие поля:

1. Название группы. Данное поле может быть использовано как название раздела предметной области, к которой относится данная группа (раздел).

2. Описание групп. Данное поле может использоваться для различных целей. Например, для описания раздела предметной области, с которым отождествлена группа; для описания причин выделения данной группы; для описания признаков, по которым определяет принадлежность понятий к данной группе.

Для того чтобы выделить некоторое **понятие** предметной области достаточно всего одного поля, которое и будет описывать это понятие. Каждое понятие может иметь несколько **определений**, которые зависят от контекста использования понятия. Каждое определение – это просто текст, который описывает это определение. Таким образом, может определить полную структуру хранения онтологии предметной области и онтологии проекта. Каждый словарь, может содержать в себе набор дочерних групп, в свою очередь, которые могут содержать подгруппы и так далее. Каждая группа может содержать набор понятий, относящихся к данной группе. Каждое понятие в свою очередь может содержать несколько определений. Поэтому в качестве структуры словарей использована древовидная структура, где корни дерева – это словари, а листы дерева – это определения понятий.

Так как о назначении и содержании «Словаря» было сказано выше, то пояснения в структуре необходимы только для «отношений» и «материализаций».

С различными «отношениями» в структуре типовой «онтологии проекта» связаны различные типы систематизации, которые можно ввести в конкретную онтологию O_{mk}. Приписывание отношениям имён и типов осуществляется субъектом Sb_i в операционной обстановке, фрагмент которой приведён на рисунке 6. Полезность типов, представленных в интерфейсной форме, была выявлена при создании и использовании «онтологий проектов» в инструментальной среде WIQA.Net. Отметим, что отобранный набор типов существенно богаче, чем те наборы типов, которые традиционно используются при создании онтологий.

Отметим, что отношениям можно не только приписать имена, выводящие на их семантику, но и прокомментировать. Такая возможность особо полезна для прагматических отношений, например, инструментального типа (связывающего объект с инстру-

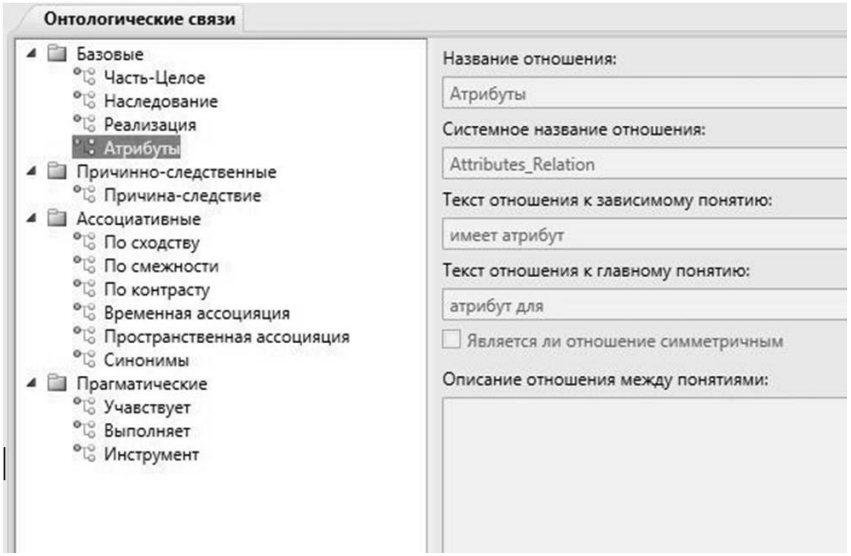

Рис. 6. Приписывание онтологических связей

ментами для его обработки или процесс со средствами его реализации).

«Материализация» введена в структуру онтологии для референциального связывания того, что стоит за понятием с «референтом», на которое понятие указывает при его употреблении в текстах или рассуждениях. Часть списка типовых «референтов» человеко-компьютерной среды, а также операционная среда их выбора и спецификации приведены на рисунке 7.

Учёт «материализации» повышает ответственность субъекта Sb_i за выбор слов, используемых им в решении задач, способствует предотвращению ошибок, а также вводит в онтологию дополнительную систематизацию. Отметим, что список «типов материализации» открыт для включения новых составляющих.

Отметим и то, что для выполнения конкретной референциальной функции необходимо представить «референт» в среде OwnWIQA определённым файлом или указать «местоположение» референта в таком файле. Спецификация референции регистрируется с помощью интерфейсной формы, представленной на рисунке 8.

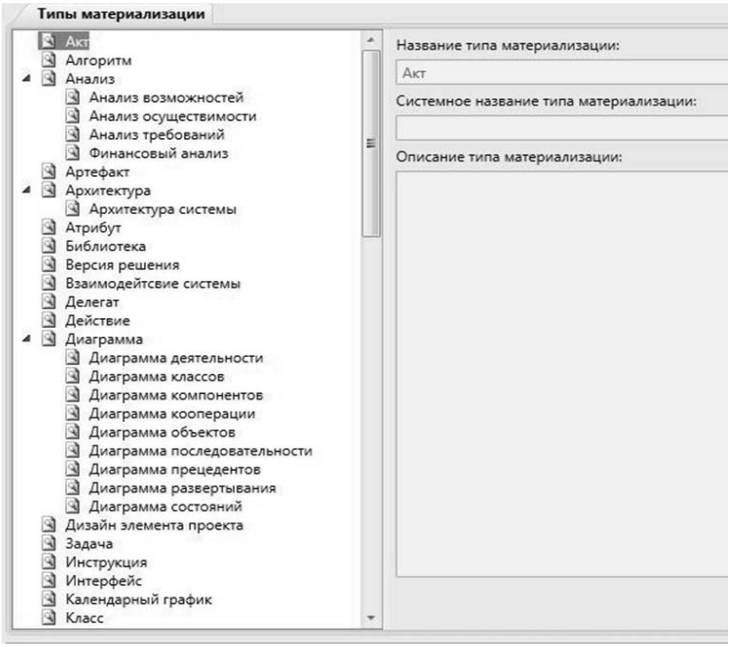

Рис. 7. Приписывание материализации

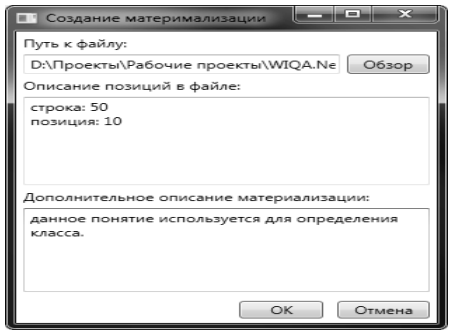

Рис. 8. Спецификация референта

Для того чтобы перейти к файлу, в котором материализовано понятие необходимо:

1. Подключить онтологию проекта, в котором находится понятие

2. В дереве онтологий выбрать необходимое понятие
3. Перейти на вкладку «Материализация»
4. В списке материализаций выбрать материализацию, к файлу которой необходимо перейти
5. Вызвать контекстное меню
6. Выбрать пункт «Перейти к файлу»

Такого рода инструкции имеются для типовых процедур OwnWIQA. В то же время для наиболее часто используемых процедур, возможно их выполнение за счёт активации определённых «горячих клавиш». Так, например, для перехода к файлу достаточно активировать клавиши Ctrl + G.

3.6. РЕАЛИЗАЦИЯ СТРУКТУР ДАННЫХ

Возможности создания и использования персональной онтологии в среде OwnWIQA обеспечиваются структурой данных, кон-

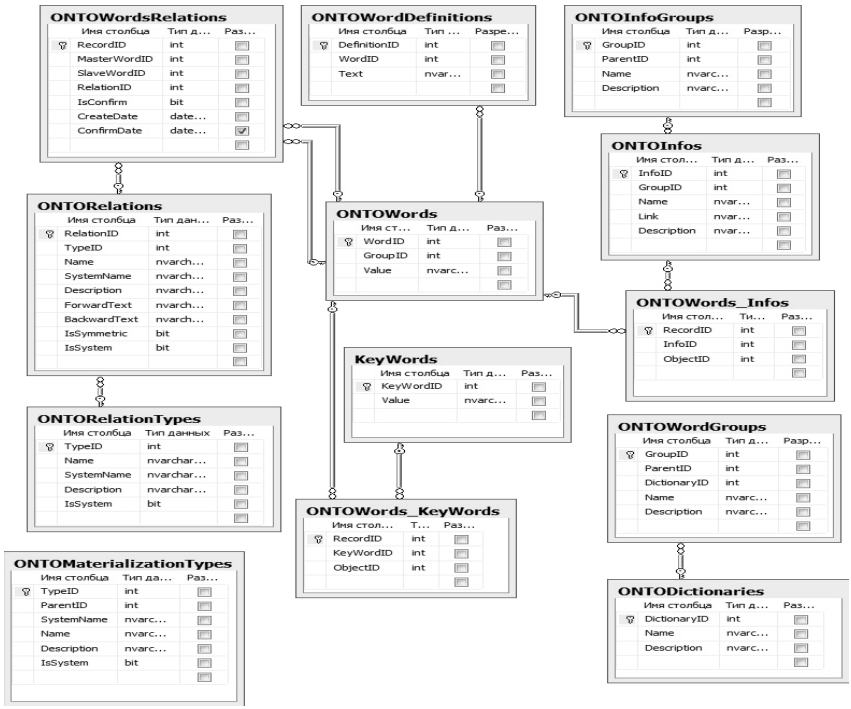

Рис. 9. Структура данных онтологии

цептуальная схема которой приведена на рисунке 9 в форме, которая реализована в двух версиях, одна из которых ориентирована на создание онтологии и её использование субъектом Sb_i, а вторая открыта для её использования в псевдо-кодовых программах.

Схема приводится с демонстрационными целями (без пояснений и спецификаций) только для того, чтобы показать масштабность информационной поддержки взаимодействий с онтологией. Её реализация по первой версии проведена по образцу программирования приложений с базами данных. Все интерфейсы, показанные выше, относятся к реализации именно этой версии.

Вторая версия представляет собой «клон» первой версии, загруженный как системный проект в QA-память комплекса OwnWIQA. Такая возможность обусловлена тем, что основу онтологии определяет иерархическая структура системы понятий. Системные отношения между понятиями представляются списками, которые регистрируются в полях описания соответствующих ячеек QA-памяти.

Подобный приём с клонированием фрагментов Базы данных используется в комплексе WIQA.Net для данных об организационной структуре коллектива проектировщиков [9].

3.7. ИСПОЛЬЗОВАНИЕ ПЕРСОНАЛЬНОЙ ОНТОЛОГИИ

В комплекс OwnWIQA, встроены два базовых варианта применения персональной онтологии, первый из которых осуществляется в традиционных формах человеко-компьютерного взаимодействия. В рамках этого варианта субъект Sb_i имеет визуальный доступ к систематизации, которую он сам построил для совокупности моделей прецедентов, использованных или используемых им в определённых проектах. В таком взаимодействии принципиальным является проверенное и контролируемое использование лексики, согласованное с пониманием прецедентов.

Второй вариант открывает дополнительные возможности, во-первых, вопросно-ответного моделирования взаимодействий с онтологией, например, для проверки корректности её содержания и связей, и, во-вторых, для псевдо-кодового программирования задач, в которых полезно применение онтологии.

К числу таких задач относится, например, поиск подходящей модели прецедента в сложившихся условиях. В OwnWIQA реше-

ние такой задачи возложено на программного агента, запрограммированного на псевдо-кодовом языке WIQA [9]. Обобщённая структура этого агента приведена на рисунке 10.

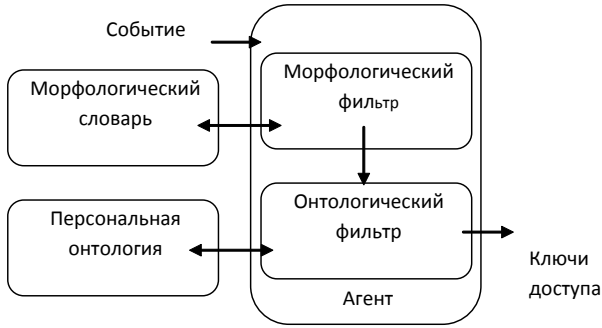

Рис. 10. Структура агента

Агент считывает с «доски поиска» описание события, затребовавшего доступ к системе моделей прецедентов. Текст этого описания обрабатывается «морфологическим фильтром», который выделяет из описания список потенциальных «ключей доступа» к Базе прецедентов. «Онтологичесики фильтр» исключает из этого списка те лексемы, которые отсутствуют в онтологии.

3.8. ИНФОРМАЦИОННЫЕ ИСТОЧНИКИ

На формирование персонального профессионального опыта принципиальное воздействие оказывают информационные источники различных типов, доступные в компьютеризованных средах. Так, например, не следует изобретать собственные определения понятий, для хорошо освоенных проблемных областей. Их следует заимствовать из тех источников, которым можно доверять.

Именно такого рода понятия следует накапливать и систематизировать в той части онтологии, которая является общей для проектов субъекта Sb_i. Следует заметить, что, заимствуя устоявшиеся понятия, следует сохранить их связь с соответствующим источником или источниками. Такая функция в OwnWIQA возложена на компоненту «Информационные источники», концептуальная структура которой приведена на рисунке 11.

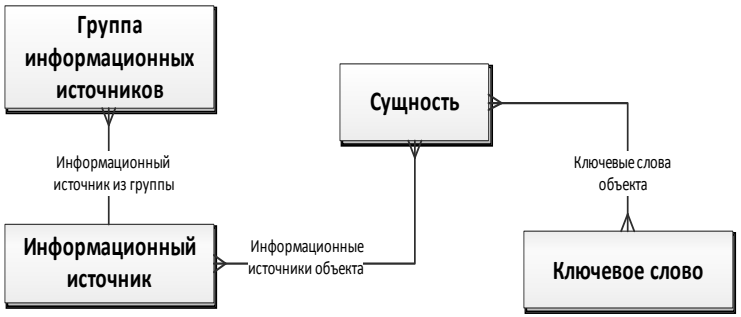

Рис. 11. Информационные источники

В концептуальной структуре присутствует блок «сущность», экземпляр которой может быть не только понятием, но и описанием определённого объекта. Независимо от того, что представляет собой экземпляр сущности, ему можно приписать «Ключевое слово» или совокупность таких слов, например для доступа в Интернет.

3.9. ПРОЕКТИРОВАНИЕ КОНФИГУРИРУЕМЫХ ШАБЛОНОВ

Для демонстрации потенциала представленных средств формирования онтологии профессионалного опыта представим их применение в профессиональной деятельности по проектированию шаблонов деталей самолётов.

В авистроительном производстве для изготовления, контроля и сбоки деталей фюзеляжа, крыльев и элерона, включая детали их обшивки, широко используется шаблонное оснащение, в состав которого входят десятки тысяч шаблонов разной степени сложности и назначения. Этот факт обусловлен такими особенностями деталей названного класса, как сложность их геометрических форм, малая ждёсткость, большие габариты, высокие требования точности изготовления и точности увязки. Кроме того, для увязки деталей, входящих в каждое плоское сечение конструкции самолёта, необходима система жестких носителей, фиксирующих контуры внутренних деталей, входящих в состав данного сечения. Так, например, большие по величине детали приходится увязывать на листовых металлических контрольно-контурных шаблонах. От-

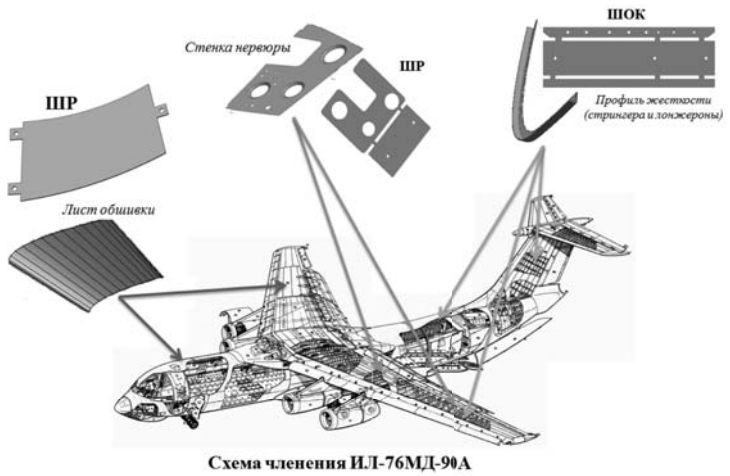

Рис. 12. Взаимосвязи шаблонов и деталей самолёта

меченные особенности неявно и обобщённо отражены на рисунке 12, на котором приведён и обозначен ряд шаблонов.

В общем случае, шаблон, не только является носителем геометрии и формы детали, но также включает конструктивные и технологические базы, контуры и оси внутренних деталей, попавших в данное сечение, различные конструктивные и технологические отверстия. Кроме того на шаблоны наносится различная информация: название шаблона, шифр и номер чертежа изделия, марка и толщина материала, указания о линиях сгиба и малки борта, контуры отверстий облегчения, маркировка отверстий и другая информация.

Представленная разнородность онтологической информации и необходимость в разработках до сотни тысяч шаблонов на очередную модель самолёта были основными причинами для разработки онтологии проектирования шаблонов на базе средств, представленных выше.

Центральное место в онтологии проектирования шаблонов, как и в любой другой онтологии, порождаемой в среде WIQA занимает «Словарь», в структуре которого выделены разделы для представления основных видов шаблонов. Статьи разделов содержат не только определения шаблонов, но также ссылки на модели шаблонов и ключи для поиска по оперативным запросам (рисунок 13).

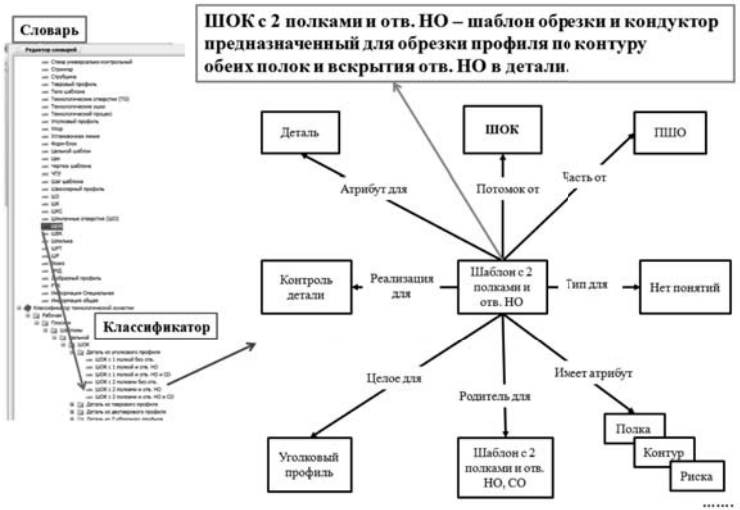

Рис. 13. Онтологический словарь: статья ШОК-Шаблон Обрезки и Кондуктор

Среда WIQA ориентирована на концептуальное моделирование в процессах решения проектных задач [....], в основу которого положено псевдо-кодовое программирование активности проектировщиков [....], что переносится и на программный доступ к составляющим онтологии, включая доступ по ссылкам к концептуальным моделям шаблонов.

Формирование классификации шаблонов предшествовало её погружению в структуры Словаря проектной онтологии, в которых сохранены особенности построения классификации для её расширения (множество шаблонов является открытым).

Благодаря инструментарию WIQA имеется возможность полного описания необходимой оснастки в онтологии словаря. Имея в наличии классифицированный шаблон и его составляющие элементы создаются связи для формализации шаблона по признакам «часть-целое», т.к. в него входят определяющие его элементы, «наследование» т.к., в большинстве своём случаев у шаблона имеется родитель или предок, и заполняются необходимые атрибуты (рисунок 14).

Отметим, что отношениям можно не только приписать имена, выводящие на их семантику, но и прокомментировать. Такая воз-

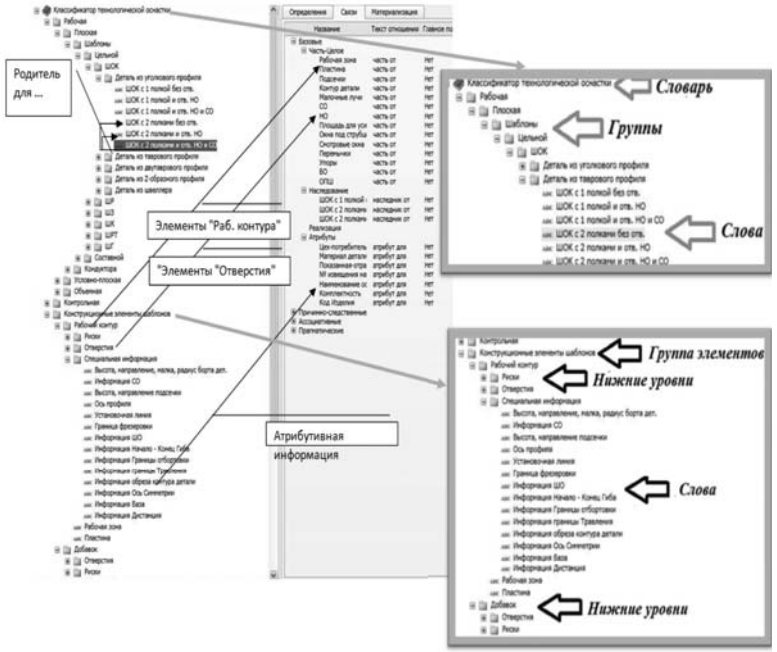

Рис. 14. Реализация Классификатора шаблонной оснастки в WIQA

можность особо полезна для прагматических отношений, например, инструментального типа (связывающего шаблон с инструментами для его обработки или процесс механической обработки со средствами его реализации).

Заключение

В работе представлены средства, предназначенные для создания и использования персональной онтологии профессионального опыта в инструментально-моделирующей среде OwnWIQA. Инструментарий ориентирован на пользователя, который (в своих собственных проектах) решает задачи, а также представляет их решения в виде моделей прецедентов для повторного использования в будущих проектах. Он же, по ходу работ, систематизируют модели прецедентов, строя персональную онтологию опыта, соответствующую освоенным прецедентам.

Спецификации персональной онтологии специально согласованы с деятельностью в рамках проектов, что позволяет: раздельно создавать онтологии проектов, объединяя их в единую систему; расширить набор типов онтологической систематизации; вводить референциальные связи составляющих онтологии с составляющими моделей опыта; использовать содержимое онтологии в программировании задач.

СПИСОК ЛИТЕРАТУРЫ

1. Данилова Н. Н., Крылова А. Л. Физиология высшей нервной деятельности / Н. Н. Данилова А. Л. Крылова. – М.: Учебная Литература, 1997.

2. Соснин П. И., Шамшев А. Б. Комплекс средств контроля семантики проектных задач и проектных решений / П.И. Соснин, А.Б. Шамшев // *Автоматизация процессов управления*. – 2010. – № 3. – С. 55–62.

3. Sosnin P. Creation and Usage of Project Ontology in Development of Software Intensive Systems / P. Sosnin // *Polibits*. – 2010. – Vol. 2. – P. 132–146.

4. Соснин П. И., Маклаев В. А. Инструментальные средства для спецификации концептуализаций в проектировании автоматизированных систем / П.И. Соснин, В.А. Маклаев // *Онтология проектирования*. – 2012. – № 2. – С. 39–52.

5. Basili V. R., Lindvall M. Costa P., Implementing the experience factory concepts as a set of experience bases / V. R. Basili, M. Lindvall, P. Costa // *Proceedings of the International Conference on Software Engineering & Knowledge Engineering*. – 2001. – P. 102–10.

6. Henninger S., Tool Support for Experience-based Software Development Methodologies / S. Henninger // *Advances in Computers*. – 2003. – Vol. 59. – P. 29–82.

7. Sjoberg D.I. K., Dyba T., Jorgensen M., The Future of Empirical Methods in Software Engineering Research / D.I. K. Sjoberg, T. Dyba, M. Jorgensen // *Proceedings of Workshop "Future of Software Engineering"*, 2007. – P. 358–378.

8. Иванов Н. В. Актуальное членение предложения в текстовом дискурсе и в языке (по материалам сопоставительного изучения португальских и русских текстов): монография. – М.: Изд. центр "Азбуковник", 2010. – 215 с.

9. Sosnin P., A Scientifically Experimental Approach to the Simulation of Designer Activity in The Conceptual Designing of Software Intensive Systems / P. Sosnin // *IEEE ACCESS*. – 2013. – Vol. 1. – P. 488–504.

Глава 4. ЛЕКСИКОГРАФИЧЕСКИЕ РЕСУРСЫ ПЕРЕВОДЧИКА: СОСТАВ, СТРУКТУРА И ВЕДЕНИЕ

Л. Н. Беляева

4.1. ВВЕДЕНИЕ

Сегодня именно деятельность переводчика оказывается базой для оперативного извлечения и дальнейшего анализа информации [1-3]. Особенно важными в этом случае являются характеристики быстроты выполнения и высокого качества перевода, поскольку перевод, выполненный поздно или некорректно, может привести к последствиям если не трагическим, то критическим, особенно в областях, связанных с реальным риском для жизни и здоровья (в медицине, сейсмостойкости, атомной энергетике и т.д.).

Современным средством поддержки работы переводчика являются информационные технологии (ИТ) и созданные с их помощью лингвистические ресурсы и системы осуществления и/или поддержки перевода. Однако, именно переводчик, воспринимающий компьютер как реальную угрозу своему существованию в профессии, часто оказывается в ситуации, когда ему не известны или недостаточно известны возможности применения информационных технологий для решения собственных задач. Незнание этих возможностей (или, что еще хуже, знание неполное и/или некорректное) приводит к тому, что переводчик не умеет оценивать и выбирать нужные именно ему средства и, следовательно, не способен их адекватно использовать. В результате, сталкиваясь с некорректным использованием ИТ и не умея получить с их помощью желаемый результат, именно переводчик часто просто отвергает саму идею использования информационных технологий в своей профессиональной деятельности.

Под лингвистическими ресурсами принято понимать естественные или искусственные языки и средства их лингвистической

поддержки, которые используются для представления информации об обрабатываемом естественном языке (словари, онтологии, тезаурусы и пр.) [4: 97-99], для представления ресурсов в системе обработки информации, для решения задач извлечения эмпирической информации. Кроме того, к таким ресурсам относятся собственно языковые источники (тексты), собранные в обширные базы данных и представляющие собой источник знаний о языках [5].

4.2. ОСОБЕННОСТИ ЛЕКСИКОГРАФИЧЕСКИХ РЕСУРСОВ ПЕРЕВОДЧИКА

Использование лингвистических ресурсов возможно как в «ручном» режиме, так и при решении задач автоматической обработки текста, в последнем случае следует учитывать:

• многовариантность результатов автоматического синтаксического анализа предложения, связанную с лексической и синтаксической омонимией, снятие которой часто вызывает затруднения даже при «ручном» анализе;

• синтаксическую и семантическую многозначность структур предложения в целом и структур именных и глагольных групп, составляющих функциональные компоненты предложения;

• особенности реализации процедур трансфера с учетом сопоставительного анализа структурных характеристик исходного языка и языка перевода (ср. [6: 120]).

Особое место в комплексе лингвистических ресурсов занимают ресурсы лексикографические, представляющие для переводчика возможность оперативного извлечения информации из различных источников: терминологических баз и банков данных, электронных словарей и словарей автоматизированных систем обработки текстов, из предметно-ориентированных корпусов текстов и т.п.

Компьютерные лексикографические (терминологические) ресурсы были хорошо приняты переводчиками с момента своего появления − с начала шестидесятых лет прошлого века. В крупных правительственных и промышленных организациях была осознана неотложная потребность в быстром доступе к современным глоссариям и словарям в области науки, техники, экономики и социальных наук в целом. Трудности были абсолютно ясными:

быстро изменяющаяся терминология многих научно-технических дисциплин, появление новых понятий, новых методов и новых продуктов, часто недостаточная стандартизация терминологии и многочисленность источников информации различного качества и надежности. По оценкам того времени переводчики могли тратить до 60% своего времени на консультации со словарями, глоссариями и другими терминологическими источниками [7].

С самого начала стало очевидно, что словари, предназначенные для переводчиков, не могли быть словарями, разработанными для систем машинного перевода (МП), как предполагалось ранее [8]. Переводчики не нуждаются в детальной информации о грамматических функциях, синтаксических категориях, семантических признаках, флективных формах и т.д., которая абсолютно необходима для автоматического анализа (парсинга). Кроме того, переводчикам не нужны словари общего языка, которые столь же необходимы в системе МП, анализирующей предложение в целом. Прежде всего, переводчику необходим доступ к специализированной технической и научной терминологии с переводами, которые являются стандартизированными и одобренными экспертами в соответствующих предметных областях.

Следует отметить, что системы, работающие с терминологией, существуют достаточно давно. Еще в 70-х годах XX века крупные компании и правительственные организации создавали машинные языковые фонды: параллельно с экономическим и техническим ростом постоянно появлялась новая терминология, и такие фонды предназначались для унификации терминов, использующихся в определенных типах текстов и при переводе. В это время одним из наиболее крупных фондов был TEAM, разработанный компанией Siemens для работы с европейскими языками, в частности, с русским, он включал около 700000 лексических единиц из различных тематических областей (естественные науки, бизнес, техника и т.п.), соответственно сгруппированных [9]. Материалы этого фонда используются и в настоящее время при создании специализированных словарей. В 80-е годы прошлого века разрабатывалась концепция Машинного фонда русского языка, в задачу которого входило создание универсальной базы данных русского языка (см., например, [10]).

В задачу терминологических ресурсов входило представление необходимой информации об отдельных словах или словосочета-

ниях (описания, примеры, переводы), эти фонды использовались в качестве основы для создания глоссариев специальных текстов и для автоматизации издания специализированных переводных словарей [11].

4.3. ТЕРМИНОЛОГИЧЕСКИЕ БАНКИ ДАННЫХ КАК СЕТЕВОЙ ЛЕКСИКОГРАФИЧЕСКИЙ РЕСУРС

В 90-х годах прошлого века с развитием вычислительной техники и появлением сетевых возможностей хранения и передачи информации начали разрабатываться терминологические базы (или банки) данных (ТБД). Эти базы представляют собой автоматизированные хранилища, в которых термины снабжаются дополнительной информацией как лингвистического (сочетаемость, частотность, принадлежность к семантическому полю), так и экстралингвистического (нормативность, стандартизованность и т. п.) типа. «В зависимости от цели создания ТБД их можно разделить на две группы: ориентированные на обеспечение работ по переводу научно-технической литературы и документации и предназначенные для обеспечения информацией о стандартизованной и рекомендованной терминологии» [12: 284].

Многие из банков данных сразу создавались как многоязычные, почти у всех предусматривался прямой доступ в диалоге, у большинства было развернутое описание единиц, являющихся заглавиями словарных статей, и некоторые из банков того времени были очень большими. В случае с терминологическими банками особое внимание уделялось условиям работы, «дружественности» интерфейса. Новые термины снабжались примерами, текстами на другом языке, дефинициями, полученными из надежных источников, кодами предметной области и библиографическими ссылками.

Терминологические ресурсы по степени универсальности и доступности можно разделить на государственные (например, поддерживаемые Комиссией ЕС) и инициативные, разрабатываемые корпорациями или исследовательскими группами.

Банк данных Eurodicautom [http: // www.mt-archive.info/LREC-2000-Johnson.pdf] представляет собой один из самых мощных государственных терминологических банков, охватывая все языки Европейского союза и латынь. В основную словарную базу к 2008

году было включено 1.240.000 словарных статей (5 миллионов терминов) и 325 000 аббревиатур и акронимов. Коды предметных областей основаны на универсальной классификации Леноха.

При создании и ведении этого банка терминов ввод информации был организован в каждом из переводческих бюро ЕС в соответствии с собственными правилами и подходами в зависимости от различных соглашений об использовании и методов сотрудничества каждого языкового сообщества и каждой страны. Целесообразно было все отдельные базы данных объединить в согласованную базу, разрешающую постоянный ввод материала приблизительно 5000 переводчиков из учреждений ЕС. Пополнение словарной базы данных происходило за счет работы терминологического бюро (в Брюсселе и Люксембурге), предложения, поступающие от переводчиков, систематизировались группой Eurodicautom, кроме того, часть информации поступала по контрактам от частных компании и экспертов в отдельных областях знаний. Обновление системы происходило еженедельно [13].

В 2008 году в Европейском парламенте было принято решение создать особую структуру, которая должна координировать исследования в области терминологии, согласовывать получаемые из различных источников данные, заниматься сохранением этих данных в формате IATE (InterActive Terminology for Europe) – лингвистического ресурса в виде терминологической реляционной базы данных, а также сотрудничать с учреждениями при ведении новой базы данных, содержащей миллионы терминов, извлеченных из других баз и импортируемых без всякой фильтрации. Ведение базы предусматривало удаление устаревших и вышедших из употребления терминов, а также дубликатов недавно добавленных единиц. В качестве такой структуры Европейским парламентом организован отдел по координации терминологии TermCoord, который осуществляет доступ к терминологии ЕС через общедоступный сайт и бесплатные инструментальные средства, а также через Межведомственный терминологический портал EurTerm [14].

Для терминологической работы в рамках отдела организована специальная группа переводчиков, включенных в терминологическую сеть, объединяющая более сотни переводчиков в 23 бюро переводов, обслуживающих разные языки, по крайней мере, по два человека на каждое бюро. Для этих переводчиков-терминологов

необходимы специальные навыки и знания. Соответственно, на первом этапе потребовалось организовать специальное обучение в области ведения базы данных, предварительное обучение для переводчиков, на этой же основе регулярно осуществляется начальная подготовка новых сотрудников и в год в среднем обучаются 60 стажеров из разных переводческих бюро, работа которых в Парламенте ограничивается 3 месяцами.

Для поддержания базы данных в актуальном состоянии в описываемом ресурсе используются различные инструментальные средства: процедуры извлечения терминов из текста, макросы для облегчения предварительного хранения терминов в процессе перевода, серверы поиска для визуального просмотра сотен связей со специализированными импортированными глоссариями, web-страницы, концентрирующие связи со всеми справочниками всех организаций ЕС. Некоторые из этих ресурсов (DocHound и GlossaryLinks) были опубликованы на сайте отдела и используются внешними подрядчиками, переводящими до 30% текстов Парламента.

Планируется, что терминологический портал EurTerm, разработанный отделом TermCoord, будет иметь доступ к внутреннему формату IATE, централизованный доступ ко всем межведомственным и мировым терминологическим ресурсам и банкам данных, а также доступ к платформам коммуникации на уровне языка (терминологии wikis и форумы). Кроме того, он будет обеспечивать доступ к инструменту QUEST для мета-поиска терминологии внутри ЕС, этот инструмент осуществляет поиск в наиболее важных национальных терминологических базах данных в дополнение к IATE, переводческой памяти Euramis и EurLex [14].

В результате проведенных исследований и организационных мероприятий в этом лингвистическом ресурсе объединено 133 локальных ресурса, разработанных в различных бюро перевода. Предусмотрены связи с 4 базами данных (базой IATE, базой данных Латвийской Академии Наук TermNet.lv, открытым терминологическим словарем сельскохозяйственной академии г. Щецин OSTEN, венгерской базой MoBiDic, поддерживающей переводы с венгерского языка). В ресурсе 2 650 976 терминов (число постоянно увеличивается), 710 705 словарных статей, 221 512 дефиниций на 33 языках [http: // www.eurotermbank.com/Collection_list.aspx?langu=bg].

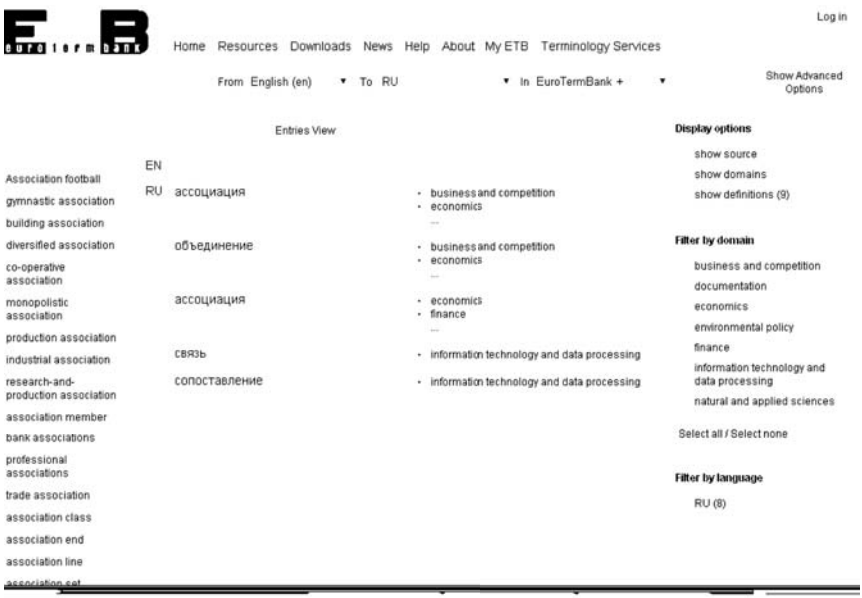

Рис. 1. Структура выдачи информации о переводе термина association

Структура информации в базе данных EuroTermBank предполагает различные опции выбора исходного языка и языка перевода, выбора предметной области, выбора формы представления информации. При выборе конкретных опций предоставляется информация о вариантах перевода в различных предметных областях и о зафиксированных в базе данных словосочетаниях. На рис. 1 и 2 представлена структура выдачи информации для перевода с английского на русский и с русского на английский. Обращение к словарю осуществляется бесплатно.

Использование ресурса в его современном виде позволяет переводчику осуществлять поиск терминов в различных источниках, идентифицировать кандидаты в термины собственных документах и автоматически извлекать их, просматривать варианты перевода термина в разных предметных областях, искать термины в нескольких языках перевода одновременно, уточнять переводы и делиться информацией с другими пользователями. Доступ к ресурсу осуществляется непосредственно из Microsoft Word.

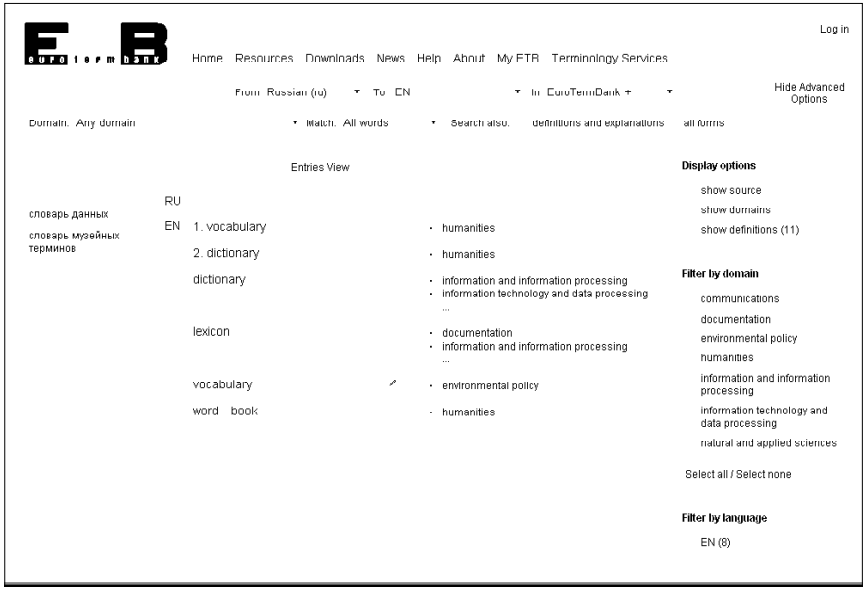

Рис. 2. Структура выдачи информации о переводе термина словарь

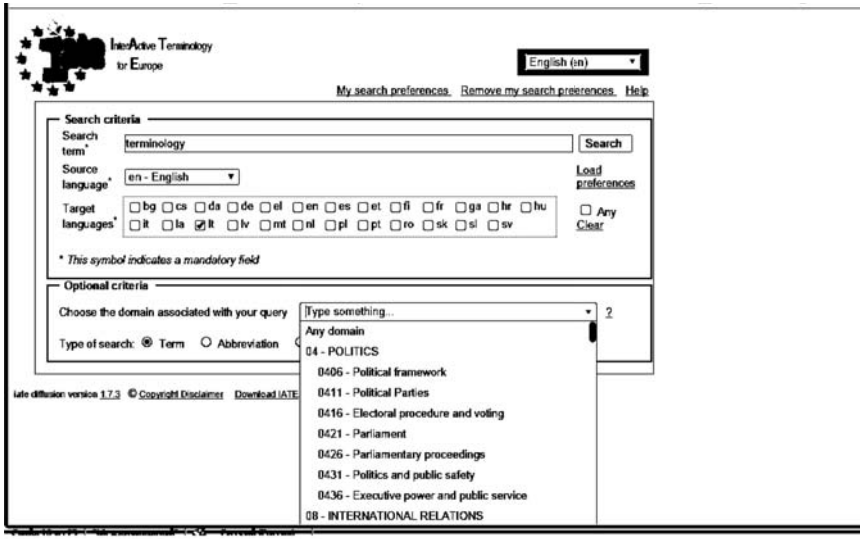

Рис. 3. Структура выбора информации в системе IATE

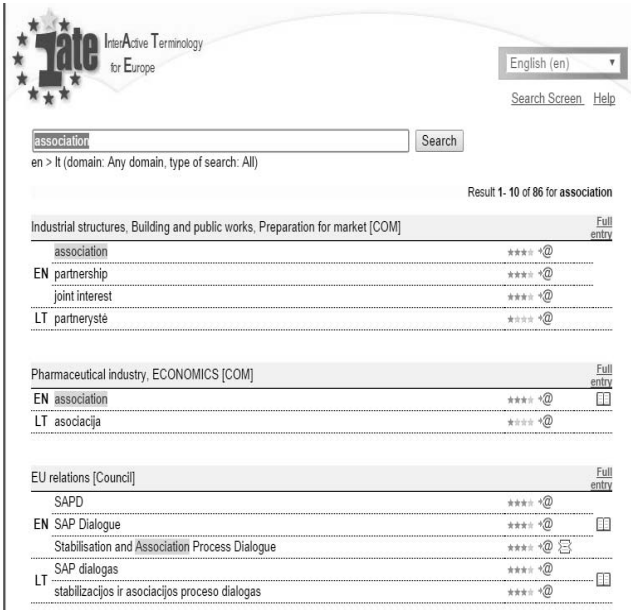

Рис. 4. Пример выдачи информации по запросу в системе IATE

В свою очередь, система IATE (InterActive Terminology for Europe), реализованная как реляционная база данных и связанная с ресурсом EuroTermBank, жестко ориентирована на языки Евросоюза и поддерживает только их (см. Рис. 3 и 4).

Информация о переводе по запросу пользователя включает возможность выбора пары языков из 25 возможных, выбора конкретной предметной области и подобласти в ней в соответствии с принятой в базе классификацией, определение типа лексической единицы – термин или аббревиатура.

Ресурс базы терминов EuroTermBank, подробно описанный выше, может рассматриваться как опробованная модель многоязычного сетевого ресурса, создание которого является безусловно актуальным как для языков национальных республик России, так и для языков таможенного союза, поскольку может обеспечить корректную терминологическую и лексикографическую поддержку для перевода документов в различных областях сотрудничества.

В то же время следует иметь в виду, что лексикографические ресурсы, подобные описанным выше, включают главным образом терминологию, извлекаемую в результате стандартизации, и (несмотря на огромные объемы) не способны охватить всю терминологию, особенно для активно развивающихся областей знаний. Сегодня основными недостатками терминологических ресурсов является высокая стоимость и длительное время, необходимое для их создания, недостаточный охват терминологии, особенно для номинации самых современных понятий, недостаточность совместного использования терминологических ресурсов и отсутствие механизмов для вовлечения терминологов-практиков. Новым «облачным» ресурсом, предназначенным для устранения этих недостатков, является разработка сервисной модели TaaS (Terminology as a Service), задачей которой является автоматизация основных этапов терминологической работы и оперативность создания и ведения [15]. Модель TaaS основана на принципе взаимности. Пользователи обрабатывают документы, и уточняют и обогащают получаемые терминологические данные, которые затем могут совместно использоваться и передаваться другим пользователям, а также вноситься в банки терминов.

Соответственно, TaaS предлагает набор функционально совместимых «облачных» сервисов, объединенных в последовательности выполнения работ:

• автоматическая идентификация кандидатов в термины в предоставляемых пользователем одноязычных документах,

• поиск эквивалентов перевода для извлеченных одноязычных кандидатов в термины;

• извлечение кандидатов в переводной эквивалент из параллельных или сопоставимых данных Интернета для терминов, отсутствующих в существующих ресурсах, с использованием методов извлечения одноязычных терминов и выравнивания пар кандидатов в термины.

При этом переводные эквиваленты извлекаются

• из терминологических банков в режиме “онлайн”,

• из автоматически извлекаемой многоязычной терминологии из сопоставимых и параллельных ресурсов Интернета,

• из коллекций терминов, созданных пользователями платформы.

Терминологические и, шире, лингвистические ресурсы используются как специалистами в области языка и перевода, так

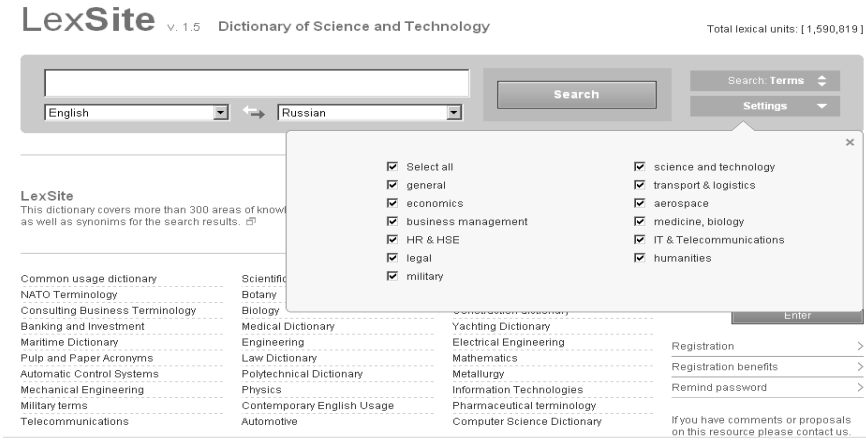

Рис. 5. Общая структура интерфейса системы LexSite

и различными программами автоматической обработки текстов, поэтому в платформу TaaS включено средство поиска терминов API для доступа подобных систем к терминологическим сервисам и данным [15]. Этот проект развивается очень активно и является перспективным лексикографическим ресурсом, ведение которого является коллективной терминологической работой всех свободно регистрируемых участников.

К корпоративным терминологическим системам можно отнести сетевой словарь LexSite [http: // www.langint.com/lexsite], который представляет собой открытый лексический ресурс, основанный на базе данных компании Language Interface. Эта база содержит общую и специальную лексику, а также уникальные термины, накопленные в результате работы над проектами в различных областях знаний. Корпус параллельных текстов, составляющий второй компонент базы данных, включает пары типа оригинал − перевод для российских нормативных и юридических документов [16]. Результат поиска ранжирован по 300 предметным областям, из которых 13 являются базовыми (см. рис. 5). Словарные статьи снабжены синонимической информацией, которая выводится на экран по запросу (см. рис. 6)

Описанные выше системы хранят заранее отобранные термины и их переводы, несколько другие функции выполняют лингви-

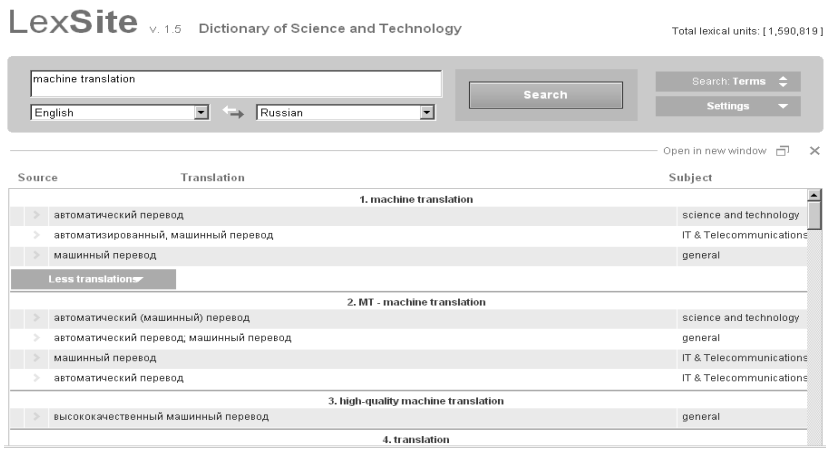

Рис. 6. Пример интерфейса системы LexSite при поиске слова

стические ресурсы работы с терминологией. Системы управления терминологией (менеджмента терминологии) поддерживают функции обнаружения, оценки и использования терминологических единиц в рамках конкретной предметной области или (уже) конкретной организации. Эти системы поддерживают деятельность авторов, экспертов, редакторов и переводчиков (см., например, систему Acrolinx IQ terminology manager [http: // www. acrolinx.com/acrolinx-iq-terminology_en.html]).

В подобных системах предполагается, что автор текста описывает новый продукт, процесс или устройство, для которых в исходном языке может не быть принятых номинаций. В этом случае с помощью глоссариев системы автор имеет возможность:

• проверить статус выбираемого термина, который может быть предлагаемым, не рекомендуемым, принятым, предпочтительным, в некоторых случаях слово имеет статус «не-термин»,

• создать новый термин на основе имеющихся онтологии или тезауруса.

В задачу эксперта в предметной области входит оценка правильности терминов относительно выбранной предметной области и языка для специальных целей, а также связанной с термином информации. Технический редактор проверяет последовательность использования терминологии во всем проекте в целом

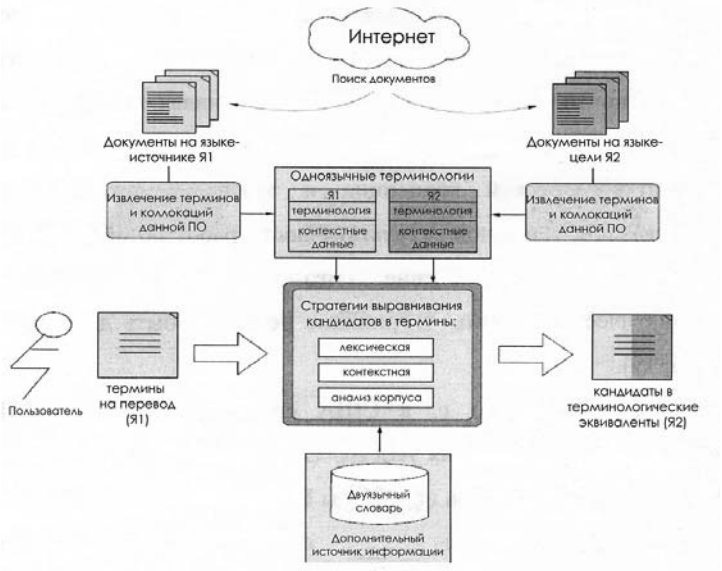

Рис. 7. Структура процесса гармонизации терминологии

и оценивает лингвистическую правильность используемых терминов.

В задачу переводчика при использовании такой системы входит:

• установление существующей в исходном языке терминологии,

• исследование значения выделенных терминов в языке перевода,

• создание новых терминов в языке перевода.

Редактор перевода также оценивает последовательность использования переводной терминологии во всем проекте в целом.

Несмотря на то, что термин является центральным элементом, интересующим и терминолога, и переводчика, работающих в конкретной предметной области, − термин является единицей словаря, объектом перевода и извлекаемым из корпуса элементом текста. При создании терминологической базы данных необходимо соотносить получаемые термины с концептуальным описанием данной предметной области, как это реализуется, на-

пример, в рамках многоязычного проекта EuroTermBank [13]. В качестве универсальных описаний в больших проектах используются, например, тезаурус Eurovoc или онтологии, включающие уже несколько тысяч понятий (ср. OMEGA, SUMO, DOLCE и др.). Подобное описание, созданное для каждого из рабочих языков, позволяет выделить ядерные понятия исследуемой области, а также оценить сбалансированность корпуса текстов, выделить зоны, требующие особого внимания в связи с появлением новых понятий [17]. Последующая гармонизация терминологий может осуществляться по схеме, реализуемой в рамках проекта Штутгартского университета TTC (рис. 7) [TTC Project http: // www.ttc-project.eu/about-ttc/concept-and-objectives].

Из текстов на исходном языке, которыми располагает переводчик или исследователь, можно извлечь начальный список слов для дальнейшего целенаправленного поиска документов, которые составят основу корпуса (или нескольких корпусов) текстов. Такая же процедура проводится и для текстов на языке перевода. Данный начальный список слов может также послужить основой терминологической базы и онтологии для конкретных предметных областей. При автоматическом выделении опорных терминоэлементов следует с осторожностью относиться к критерию частотности, так как в верхней части списка могут оказаться не только строевые элементы текста, но и общенаучные или общетехнические лексические единицы, «ненужность» которых довольно сложно заранее вычислить и указать в стоп-листе.

Для выравнивания пар терминов на двух языках в проекте TTC используется комбинация статистических и лингвистических методов, с помощью которых лексические единицы, извлеченные из текстов на обоих языках, группируются в эквивалентные пары. При этом этапы Интернет-поиска и извлечения терминов полностью автоматизированы, а последующие этапы автоматизированы частично [TTC Project http: // www.ttc-project.eu/about-ttc/concept-and-objectives].

Проект «Интертекст» [18-19] – многоязычная база знаний, для создания которой необходимы «семантико-синтаксический анализ и системное моделирование межъязыковых соответствий когнитивных структур полнотекстовых научных и патентных документов для задач обработки знаний и машинного перевода [18: 177].

Наиболее популярными и наиболее мощными англо-русскими сетевыми словарями являются Мультитран, ABBYLingvo и Мультилекс [16: 151-153], причем Мультитран превосходит остальные ресурсы по наполнению, но уступает им в быстродействии [16: 154].

4.4. АВТОМАТИЗАЦИЯ ПРОЦЕДУРЫ ИЗВЛЕЧЕНИЯ ТЕРМИНОВ ИЗ ТЕКСТОВ НА РАЗНЫХ ЯЗЫКАХ

Идея создания автоматизированных систем извлечения терминов из корпусов параллельных текстов насчитывает уже более 20 лет и в той или иной степени реализуется в различных проектах. Очень важно понимать, что даже самая изощренная система извлечения терминов не дает окончательного результата для включения в переводной словарь, а предоставляет лишь удобно организованный и оперативно получаемый ресурс для работы терминолога или лексикографа. Методы извлечения терминов варьируются от независимого от конкретного языка извлечения n-граммов с использованием оценки относительной частоты и степени терминологичности [20-21] до лингвистически обоснованных методов на основе синтаксического анализа и применения моделей терминологических словосочетаний [22]. Комбинация статистических и лингвистически обоснованных приемов [23-25] наиболее используемым подходом в практических инструментальных средствах создания и лексикографических ресурсов.

С точки зрения исходной информации существуют два основных подхода к проблеме автоматического извлечения из корпусов текстов информации о переводных соответствиях двух языков и построения на ее основе лексических конкордансов. Решение задачи в рамках первого подхода начинается с выравнивания параллельных текстов. Для выравнивания параллельных текстов обычно используются определенные эвристические соображения (или просто эвристики), которые помогают выбрать точки соответствия (точнее кандидаты на то, чтобы ими быть). В роли последних могут выступать, например, числа, аббревиатуры, даты, имена собственные, акронимы, устойчивые словосочетания, имеющие однозначные переводные эквиваленты на втором языке. Для достаточно близких языков, как, например, испанский и пор-

тугальский в роли пар кандидатов могут рассматриваться слова, имеющие одинаковые корни.

Если в более ранних работах эвристики использовались напрямую, без какой-либо поддержки статистическими методами, то в настоящее время использование статистических методов играет ключевую роль в содержании большинства публикаций по проблеме автоматического выравнивания параллельных текстов. Наиболее общий метод статистической поддержки, инвариантный относительно того, какие эвристики были использованы при создании исходного списка пар-кандидатов, основан на использовании корреляционной зависимости между позициями точек (лексических единиц), составляющих пару [17].

На этом пути кроме сложностей, связанных с отсутствием симметричности терминологических систем разных языков, особую проблему представляет отбор переводов для корпуса параллельных текстов, поскольку их качество часто является сомнительным. Поэтому обращение к сопоставимым корпусам текстов, при организации которых возможна экспертная оценка текстов на сопоставляемых языках, вполне естественно.

Следует учитывать, что при использовании корпусов сопоставимых текстов вопрос о выравнивании переходит в особую плоскость. В случае параллельных корпусов текстов основным является выравнивание по предложениям, которое опирается на формальные показатели границ и частей предложений, соответствие объемно-прагматических структур текстов. При всех возникающих технических и лингвистических сложностях этот процесс вполне реализуем. В случае текстов сопоставимых возможно только терминологическое выравнивание, опирающееся на выявление характерных для обоих массивов корпуса однословных терминологических единиц и их сопоставление в качестве кандидатов в переводные эквиваленты, а также поиск устойчивых словосочетаний с этими однословными терминами в качестве ядер [26]. Дальнейший сопоставительный анализ требует привлечения знаний из переводных автоматизированных словарей, позволяющих верифицировать выбранные пары терминов. Извлечение многокомпонентных терминов может производиться на основе автоматического синтаксического анализа на уровне функциональных сегментов – именных групп.

Так, например, выявление кандидатов из корпусов однословных терминов может опираться на семантические характеристики этих слов, извлекаемые из различных автоматизированных баз данных. При использовании в качестве справочного массива словарей предметно-ориентированных систем машинного перевода, словарные статьи которых содержат базовые синтаксические и семантические характеристики выявленных слов, такое соотнесение можно автоматизировать.

Таким образом, автоматическое извлечение терминов (как универбов, так и многокомпонентных лексических единиц – коллокаций) основано на предварительном выравнивании текстов на разных языках, идентификации терминологических единиц в текстах на одном языке и дальнейшем установлении их переводных эквивалентов или, скорее, кандидатов в возможные переводные эквиваленты. Хотя утверждается, что подобная задача хорошо решена для разных языковых пар в случае анализа параллельных текстов, ее решение еще требует исследований в случае сопоставления языков с различной графикой [15].

Итак, большинство автоматизированных систем извлечения терминов используют либо статистический, либо лингвистический подход [27-28]. При этом используется частота лексической единицы в тексте, отношение правдоподобия для двусловных терминов, мера, основанная на полном количестве информации. Для оценки коллокаций, состоящих более чем из двух слов, в качестве единственного статистического параметра используется частота кандидата в термины в корпусе текстов [25].

В последнее время появились гибридные подходы, использование которых представляет собой попытку преодоления ограничений односторонних подходов к решению задачи извлечения терминов на основе как лингвистических, так и статистических элементов [29].

Одним из методов оценки степени терминологичности является независящий от предметной области метод автоматического выявления многокомпонентных терминов в тексте. В качестве исходного материала для анализа при этом используется корпус текстов исходного языка, на основе которого формируется список кандидатов в многокомпонентные термины. Эти термины упорядочиваются по степени терминологичности, которую принято называть C-value. Получаемый в результате список оценивается

экспертом в конкретной предметной области. Поскольку кандидаты в многокомпонентные термины оцениваются по степени их терминологичности, эксперт может просматривать списки, начиная от верхней части сверху вниз, работая с ним столько, сколько позволяет время и/или затраты [25].

Подход с использованием C-value основан на объединении лингвистической и статистической информации, причем особое значение имеет именно статистическая компонента. Лингвистическая информация состоит из грамматической разметки корпуса текстов по частям речи, лингвистический фильтр ограничивает тип извлекаемых терминов и включает список стоп-слов (в другой терминологии – антипризнаков).

Лингвистическая база, необходимая для реализации этого метода, включает следующие компоненты:

1. Информацию о части речи, извлекаемую из результатов парсинга и грамматической разметки корпуса текстов.

2. Собственно лингвистический фильтр, применяемый к размеченному корпусу текстов, чтобы исключить те цепочки, извлечение которых не требуется по формальным признакам. К моделям таких цепочек относится неразрешенная комбинаторика частей речи. При этом возможно применение как «закрытого» фильтра, разрешающего извлечение цепочек слов только конкретных типов, так и «открытого» фильтра, в котором перечисляются только неразрешенные типы цепочек.

3. Список слов-антипризнаков [26].

В рамках собственно статистического анализа исследуются статистические характеристики цепочки лексических единиц, являющихся кандидатами в термины. Необходимость гибридного подхода определяется тем, что доступная для анализа статистическая информация без специальной лингвистической фильтрации не является достаточной для получения достоверных и/или полезных результатов. Так, например, без учета лингвистической информации бессмысленные в терминологичском смысле цепочки слов типа *is a* также будут извлекаться.

Выбор конкретного лингвистического фильтра и его наполнение зависят от того, как терминолог предпочитает сбалансировать полноту и точность: предпочтение точности над полнотой, вероятно, потребует использовать закрытый фильтр, в то время

как предпочтение полноты определяет использование открытого фильтра [25].

Мера C-value строится достаточно прямолинейно на основе характеристик цепочек лексических единиц, являющихся кандидатами в термины. К таким характеристикам относятся:

1. Суммарная частота цепочки лексических единиц, являющихся кандидатами в термины, в корпусе текстов.

2. Частота цепочки лексических единиц, являющихся кандидатами в термины, как часть других более длинных кандидатов в термины.

3. Количество таких более длинных кандидатов в термины.

4. Длина цепочки лексических единиц, являющихся кандидатами в термины (в количестве слов).

Значение C-value вычисляется в зависимости от длины коллокаций, начиная с самых длинных и заканчивая биграммами.

Большинство лингвистических подходов основано на использовании синтаксических моделей и систем фильтров. Как правило, термины описываются регулярным выражением из меток частей речи, извлекаемых на основе анализа последовательности слов текста. Примером такой системы является система TERMS [23], аналогично работает и система LEXTER. Несмотря на провозглашаемый лингвистический подход, обе системы используют и некоторую базовую статистическую информацию.

Методы, используемые в статистических системах, варьируются от простых подсчетов частот до вычисления сложных статистических индикаторов для измерения силы связи элементов коллокаций, встретившихся в структуре кандидата на роль термина. Основные проблемы, возникающие при применении этих подходов, состоят в том, что частые слова или сочетания слов с высоким индексом связи не обязательно являются терминами. В некоторых статистических подходах привлекаются лингвистические данные, которые должны позволить преодолеть эти ограничения.

Слова и словосочетания, извлеченные подобными системами из предметно-ориентированных корпусов текстов, не всегда релевантны с терминологической точки зрения, хотя могут быть лексическими единицами, зафиксированными в этих корпусах, поэтому их принято называть *term candidates* – кандидатами в термины (КТ) [17].

Чтобы установить, какие кандидаты действительно представляют собой единицы перевода, то есть для данного типа исследований термины, а какие должны быть отброшены, часто используется методика определения веса термина, позволяющая количественно определить потенциал КТ быть реальным термином. Чаще всего для определения веса ЛЕ и оценки степени ее терминологичности используется подход, основанный на сравнении корпусов. Этот подход, который принято называть Contrastive Automatic Term Extraction [30; 31] представляет собой процедуру автоматического извлечения терминов из сопоставляемых корпусов текстов (контрастивное автоматическое извлечение терминов – КАИТ), подход основан на сравнении частот кандидатов в термины в двух различных корпусах текстов (общий корпус текстов и специализированный корпус текстов), позволяющим оценить то, что рассматривается как «норма» и фиксируется в национальных корпусах текстов. Количественный анализ этого отклонения и используется как показатель степени терминологичности.

Методы, используемые для извлечения терминов, дают возможность реализовать соответствующую структуру для объединения ряда приемов, обычно используемых для решения этой задачи. Фактически, комбинация множества классификаторов является именно той методикой, которая успешно применяется в ряде задач обработки текстов на естественном языке, например, в задачах грамматической разметки, синтаксического анализа, или классификации и фильтрации текстов, приводя к существенному улучшению результатов, получаемых на основе отдельно применяемых методов [29: 516].

Для выделения многокомпонентных терминологическх единиц разработано более 80 различных метрик. Следует отметить, что в принципе все используемые сегодня статистические оценки терминологичности и синтагматической устойчивости, по сути, являются эвристиками, поскольку в предлагаемые формулы вводятся коэффициенты, которые позволяют получить корректные результаты для различных предметных областей. Необходимость таких эвристик связана с тем, что (как показывают исследования) в различных языках для специальных целей структуры терминологических сочетаний различаются кардинально.

Лексикографические ресурсы могут быть ориентированы на выбор и обработку терминов и/или понятий, на работу с конкретной языковой парой, многоязычными или одноязычными ресурсами. Подобные средства целесообразно включать в автоматизированное рабочее место лексикографа, на основе которого можно осуществлять запись, обработку, сохранение и использование различных лингвистических и лексикографических данных.

4.5. ЛЕКСИКОГРАФИЧЕСКИЙ РЕСУРС В СОСТАВЕ ОБРАЗОВАТЕЛЬНОЙ СРЕДЫ ВУЗА

Лексикографические ресурсы – собственно терминологические базы и банки данных и системы извлечения и описания терминологии создаются и используются не только в исследовательских проектах и в технологических системах обработки информации, одной из функций которых является поддержка профессионального перевода, но и в рамках создания и ведения лингвистического обеспечения образовательной среды вуза. При этом необходимо различать возможности работы с лингвистическими ресурсами отдельных пользователей, то есть их самостоятельной работы, экспертов и менеджеров общего лингвистического ресурса, формируемого на основе словарей, создаваемых в рамках индивидуальных автоматизированных рабочих мест (АРМ) [32].

В АРМ, предназначенном для работы терминолога и лексикографа, должна быть предусмотрена возможность использования заранее выбранной системы управления контентом (*content management system*), обеспечивающей хранение данных. Выбор такой системы, составляющей важную часть высокотехнологичной образовательной среды, должен осуществляться ее менеджерами. Кроме того, АРМ должен иметь доступ к онлайновым средствам работы с терминологией. В каждом индивидуальном АРМ должны быть собственные ресурсы, используемые одновременно с онлайновыми.

В системе, объединяющей индивидуальные АРМ, должно быть ясно определено, кто именно имеет доступ к словарным базам данных и на каких условиях, поскольку терминологические единицы создаются и переводятся разными пользователями

и на разных этапах формирования словарных баз. При создании общего терминологического ресурса в такой среде структуру возможностей разных пользователей, работа которых приводит к изменению баз терминов и переводных словарей в результате извлечения и анализа терминологии, можно представить на основе следующих потенциальных ситуаций [33]:

• пользователь терминологической базы данных обнаруживает в тексте термин, отсутствующий в словарной базе.

Предлагаемая пользователем пара термин-перевод может быть введена в ресурс его индивидуального АРМ, кроме того она поступает в виде запроса менеджеру баз терминологических данных, который передает ее экспертам и в дальнейшем принимается решение о введении термина в конкретный словарный ресурс.

• разработчик конкретного терминологического ресурса вводит в базу новый термин и его перевод.

В этом случае также необходимо решение о введении термина в единую базу, формируемую в образовательной среде.

• разработчик продукта любого типа вводит новый термин, который должен быть одобрен экспертами.

В качестве такого продукта в рамках вуза могут быть номинации созданных приборов, химических веществ, программных продуктов, терминология, заявленная в учебниках и учебных пособиях.

Таким образом, можно утверждать, что ресурсная база АРМ лексикографа/терминолога, оставаясь индивидуальным пользовательским инструментом, должна постоянно оцениваться с точки зрения формирования общего ресурса, к которому в рамках высокотехнологичной образовательной среды должен поддерживаться постоянный доступ со всех компьютеров соответствующей локальной сети.

Все ресурсы образовательной среды можно разделить на информационные и технологические (программные). Такое деление ресурсов на виды не зависит от типа пользователей, их принадлежности к гуманитарной или естественно-научной сферам образования и науки, не зависит от конкретных решаемых ими задач. Предметная ориентированность информационных ресурсов определяет целесообразность их иерархизации, т.е. выделения терминальных узлов (баз) и узлов высокого уровня (универсальных). При этом терминальные узлы (базы или корпуса) будут со-

ответствовать реализуемым в вузе специализациям образования, т.е. достаточно узким и релевантным именно для конкретной образовательной среды областям знаний. При этом «узость» должна определяться не связью с конкретным факультетом, а учебной и научной спецификой.

Иерархическая структура информационных и технологических ресурсов предполагает, что на верхних уровнях (уровнях общего доступа) должны быть максимально универсальные ресурсы, использование которых не связано со специализацией исследования и обучения. Эта же иерархическая структура должна реализоваться для каждого конкретного (терминального) ресурса. Так, например, при единой системе машинного перевода, ориентированной на задачи обучения и перевода, выделяются конкретные системы автоматических словарей, соответствующие областям знаний. В предметной области «Филология» общим словарем верхнего уровня является словарь общефилологических терминов; на следующем уровне выделяются подобласти лингвистики, литературоведения и образовательных технологий в филологии, для каждой из которых должен формироваться общий словарь и система предметно-ориентированных словарей (фонетика, лексикология, лексикография, морфология и т.д. для предметной области «лингвистика»). Эту систему, которая требует совместной поддержки, для решения своих задач смогут использовать специалисты и студенты разных гуманитарных факультетов – иностранных языков, филологического, социальных наук, философии, института детства, лингвистического центра.

Аналогичные системы должны будут создаваться и для других специализаций, при этом программное обеспечение систем машинного перевода, информационного поиска, лексикографических систем и систем обучения остается максимально универсальным. Выбор и использование в рамках АРМ систем машинного перевода представляют собой особую задачу, поскольку обеспечивают оперативный поиск и обработку информации.

СПИСОК ЛИТЕРАТУРЫ

1. Климзо Б. Н. Ремесло технического переводчика. Об английском языке, переводе и переводчиках научно-технической литературы. 2-е изд., переработанное и дополненное. – М.: Р. Валент, 2006. – 508 с.

2. Кривых Л. Д., Рябичкина Г. В., Смирнова О. Б. Технический перевод. – М.: Форум – Инфра-М, 2008. – 184 с.

3. Сальмон Л. Теория перевода. История, наука, профессия. – СПб – Астана, 2007. – 271 с.

4. Большакова Е. И., Клышинский Э. С., Ландэ Д. В., Носков А. А., Пескова О. В., Ягунова Е. В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. – М., МИЭМ, 2011.

5. Рычкова Л.В. Языковые ресурсы: традиции и инновации // *Материалы V Международной научной конференции «Прикладная лингвистика в науке и образовании» памяти Р. Г. Пиотровского, 25–26 марта 2010*. – СПб: Лема. – С. 306–311.

6. Козеренко Е. Б., Лунева Н. В., Морозова Ю. И., Ермаков П. В. Проектирование многоязычного лингвистического ресурса для систем машинного перевода и обработки знаний // *Системы и средства информатики*. Вып. 19. – М.: «Наука», 2009. – С. 119–141.

7. Krollmann, F., Schuck, H.J. and Winkler, U. Herstellung textbezogener Fachwortlisten mit einem Digitalrechner: ein Verfahren der automatischen Übersetzungshilfe [Production of Text-Related Term Lists with a Digital Computer: An Experiment in Computer-Aided Translation] // *Beiträge zur Sprachkunde und Informationsverarbeitung*. – 1965. – Vol. 5. – P. 7–30.

8. Oettinger, A.G. *Automatic language translation: lexical and technical aspects, with particular reference to Russian*. – Harvard University Press, Cambridge, Mass. 1960. – 380 pp.

9. Hutchins J. Machine Translation and Human Translation: in Competition are in Complementation? / J. Hutchins // *Machine Translation: Theory & Practice*. – New Delhi, 2001. – P. 5–20.

10. Лингвистическая концепция терминологического банка данных Машинного фонда русского языка. Проект / под ред. А. С. Герда. − М.: Наука, 1989

11. Hutchins J. The Origins of the Translator's Workstation // *Machine Translation*. – 1998. – Vol. 13. – P. 287–307.

12. Лейчик В. М. Прикладное терминоведение и его направления // Прикладное языкознание: Учеб. / Под ред. А. С. Герда. – СПб.: Изд-во С.-Петерб. ун-та, 1996. – С. 276–286.

13. Towards Consolidation of European Terminology Resources. Experience Recommendations from EuroTermBank Project. – Edited by: Signe Rirdance, Andrejs Vasiljevs. Riga: Tilde, 2006. – 123 p.

14. Maslias R. Combining EU Terminology with Communication and Ontology Research // Terminology and Knowledge Engineering 2014: Proceedings of the Conference, 19-21 Jun 2014. Berlin, 2014. – P. 48–56.

15. Vasiljevs A., Pinnis M., Gornostay T. Service model for semi-automatic generation of multilingual terminology resources // Terminology and

Knowledge Engineering 2014: Proceedings of the Conference, 19–21 Jun 2014. Berlin, 2014. – P. 67–76.

16. Кит М.С. О стратегии построения высокоэффективных сетевых словарей (на базе разработки словаря LexSite) // *Вестник РГГУ*. – 2010. – № 9. – С. 149–160.

17. Беляева Л. Н., Данилова О. А., Джепа Т. Л., Камшилова О. Н., Карнуп Е. В., Нымм В. Р., Чумилкин С. В. Лексикографический потенциал современных лингвистических технологий: монография. – СПб: ООО «Книжный дом», 2014. – 168 с.

18. Галина И. В. Разработка формализованных функционально-синонимических моделей некоторых именных французских и русских языковых структур // *Системы и средства информатики.* Дополнительный выпуск. – М.: Институт проблем информатики РАН, 2008. – С. 176–194.

19. Kozerenko E.B. INTERTEXT: A Multilingual Knowledge Base for Machine Translation // *Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications*. – Las Vegas, USA: CSREA Press, 2007. – P. 238–243.

20. Delač, D., Krleža, Z., Šnajder, J., Bašić, B. D., and Šarić, F. TermeX: A Tool for Collocation Extraction // *Computational Linguistics and Intelligent Text Processing. Proceedings of CICLing-2009.* – Mexico City, Mexico, 2009. – P. 149–157.

21. Pantel, P., Lin, D. A statistical corpus-based term extractor // *Advances in Artificial Intelligence.* – Springer Berlin Heidelberg, 2001. – P. 36–46.

22. Bourigault, D. Surface grammatical analysis for the extraction of terminological noun phrases // *Proceedings of COLING 1992.* – 1992. –Vol. 3. – P. 977–981.

23. Justeson, J., Katz, S.: Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text // *Natural Language Engineering*. – 1995. – Vol. 1. – P. 9–27.

24. Dagan, I., and Church, K. Termight: Identifying and translating technical terminology // *Proceedings of ANLP 1994*. –1994. – P. 34–40.

25. Frantzi K. T., Ananiadou S., Tsujii J. The C-value/NC-value Method of Automatic Recognition for Multi-word Terms // *Proceedings of Second European Conference, ECDL'98* Heraklion, Crete, Greece September 21–23, 1998. – LNCS 1513. – Berlin Heidelberg: Springer-Verlag, 1998. – P. 585–604.

26. Беляева Л. Н. Корпусные технологии извлечения терминологии в задачах переводной лексикографии // *Структурная и прикладная лингвистика*. Межвузовский сборник. Вып. 10. – СПб: Изд-во С.-Петербургского университета, 2014. – С. 169–181.

27. Kageura, K. and Umino, B. Methods for Automatic Term Recognition: A Review // *Terminology.* – 1996. – Vol. 3(2) – P. 259–289.

28. Morin, E., Daille, B., Takeuchi, K. and Kageura K. Bilingual Terminology Mining – Using Brain, not brawn comparable corpora // *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)* Prague, Czech Republic, 2007. – P. 664–671.

29. Vivaldi J., Márquez L., Rodríguez H. Improving Term Extraction by System Combination Using Boosting // *Machine Learning ECML 2001* / L. De Raedt and P. Flach (eds.). – Series: Lecture Notes in Computer Science. – Vol. 2167. – Berlin Heidelberg Springer-Verlag, 2001. – P. 515–526.

30. Gillam L. and Ahmad K. Pattern Mining Across Domain-Specific Text Collections // *Machine Learning and Data Mining in Pattern Recognition. Proceedings of 4th International Conference, MLDM 2005*, Leipzig, Germany, July 9–11, 2005. Series: Lecture Notes in Computer Science. – Berlin/Heidelberg: Springer, 2005. – Vol. 3587. – P. 570–579.

31. Drouin, P. and Doll, F. Quantifying TH through Corpus Comparison // *Managing Ontologies and Lexical Resources. Proceedings from TKE 2008.* – Copenhagen, 2008. – P. 191-206.

32. Беляева Л. Н. Лингвистические ресурсы информационной образовательной среды: состав, структура, функции // *Известия РГПУ им. А. И. Герцена: научный журнал.* – СПб, 2014. – С. 47–52.

33. Großjean, A. Corporate Terminology Management: An approach in theory and practice. –VDM Publishing, 2009. – 100 pp.

Глава 5. ОБЪЕКТНО-ОРИЕНТИРОВАННЫЙ АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТОВ

Н. В. Лукашевич

5.1. ВВЕДЕНИЕ

Задача анализа тональности, т.е. выявление мнения автора текста по поводу предмета, обсуждаемого в тексте, является одной из активно развиваемых технологий в сфере автоматической обработки текстов в последнее десятилетие. Актуальность этого приложения во многом связана с развитием социальных сетей, онлайновых рекомендательных сервисов, содержащих большое количество мнений людей по разным вопросам, в частности о разных товарах, услугах.

Задачей первых подходов к анализу тональности текстов было определить общую тональность документа или его фрагмента [1]. Такой уровень анализа предполагает, что каждый документ выражает единое мнение по поводу некоторой единичной сущности, как например в отзыве о некотором товаре.

Поскольку в документе может быть выражена разная тональность по отношению к разным упомянутым в нем сущностям, то на следующем этапе стали решаться задачи анализа тональности по отношению к заданным сущностям, упомянутым в тексте [2, 3].

Наконец, еще более детальным уровнем анализа тональности текстов является анализ мнения по конкретным свойствам или частям (так называемым аспектам) сущности, по которым автор текста может высказывать разную тональность мнения [4-8].

В [5, 10] *мнение* определяется как пятерка (e_i, a_{ij}, s_{ijkl}, h_k, t_l), где e_i – это сущность, к которой относится мнение, a_{ij} – это *аспект* (часть или характеристика) сущности, s_{ijkl}, – это *тональность* мнения относительно этой сущности и данного аспекта, h_k – это автор мнения, t_l – это время, в которое мнение высказано. При этом

мнение s_{ijkl} – может быть *положительным*, *отрицательным* или *нейтральным,* и может выражаться с разной степенью интенсивности, измеряемой, например, по шкале 1-5.

Аспекты могут быть сгруппированы в категории (далее **аспектные категории).** Для ресторанов – это обычно кухня, обслуживание, интерьер (обстановка). Также в текстах отзывов можно встретить оценку объекта в целом – *прекрасный ресторан.* Эту категорию модно рассматривать как аспектную (**аспект Объект_в_целом).** Слова и выражения, посредством которых можно сослаться в тексте на аспект сущности, называются **аспектными терминами**.

В данной статье будут рассмотрены подходы к анализу тональности текстов по аспектам. Во втором разделе мы рассмотрим подходы к классификации аспектных терминов. В третьем разделе представлены подходы к автоматическому извлечению аспектных терминов из текстов. В четвертом разделе обсуждаются подходы к автоматическому определению тональности по отношению к заданным аспектам (аспектным категориям, аспектным терминам). В пятом разделе мы рассмотрим открытые тестирования систем анализа тональности и их результаты.

5.2. КЛАССИФИКАЦИЯ АСПЕКТНЫХ ТЕРМИНОВ

Аспектные термины в предметной области могут быть классифицированы по нескольким основаниям.

Наиболее частым видом аспектных терминов являются **явные аспектные термины**, которые явно называют объект, его части или характеристики, которые оцениваются автором текста, например, *суп*, *обслуживание*, *зал* в отзывах о ресторанах.

Явные аспектные термины чаще всего выражаются существительными или группами существительного, но некоторые аспекты могут выражаться и глаголами, например, *встретить* (*хорошо*, *неприветливо*), *ждать* (*слишком долго*, *не пришлось*) при оценке качества сервиса в ресторанах.

Вторым видом аспектных терминов являются так называемые **неявные аспектные термины,** которые представляют собой слова с явно выраженным оценочным компонентом значения, которые одновременно указывают и на обсуждаемый аспект (обычно достаточно обобщенную аспектную категорию), напри-

мер, *вкусный* (*положительный+еда* в отзывах о ресторанах), *комфортный* (*положительный+комфорт* в отзывах об автомобилях). Как и другие оценочные слова, неявные аспектные термины могут сочетаться с т.н. оценочными операторами, которые меняют или усиливают их оценку: *не очень вкусный*, *не слишком комфортный*. Важность таких аспектных терминов для словарей автоматических систем анализа тональности заключается в том, что в ситуациях нераспознавания упомянутых автором эксплицитных терминов (из-за опечаток, новой лексики, сложной референции) неявные аспектные термины дают возможность извлечь позицию пользователя по отношению к некоторой аспектной категории.

Третьим видом выражения своего мнения по поводу некоторой характеристики заданной сущности является сообщение некоторого произошедшего негативного или позитивного факта, который одновременно указывает и на аспектную категорию, так и на его оценку пользователем (далее *тональные факты*).

Одним из видов тональных фактов являются технические проблемы, упоминаемые в отзывах [12-14]. В [12] указывается, что упоминание технических проблем часто включает в себя:

– набор специального вида глаголов, обозначающих, что что-то случилось (*fail, crash, overload, trip, fix, mess, break, overcharge, disrupt*);

– набор глаголов, обозначающих, что что-то не случилось, и часто эти глаголы упоминаются с отрицаниями, а также с глаголами операторами вида (*stop*, *refuse, cease* – прекратить, прекратиться, остановиться и др.),

– некоторыми глаголами с частицами (*knock off, knock out, hang up*),

– а также существительными и словосочетаниями.

Вместе с тем тональные факты могут включать и значительно более широкий спектр ситуаций, чем технические проблемы, как например обнаружение чего-то нежелательного: «Два раза был в этом ресторане, и оба раза **нашел в своей тарелке волос»**. Liu [5] приводит следующий пример тонального факта: «*I bought the mattress a week ago, and a **valley has formed**»* (*«Я купил матрас неделю назад, и **уже образовалась впадина**»*).

Близкие по смыслу тональные факты могут выражаться в тексте разнообразными способами, что затрудняет их обнаружение.

Однако частым признаком такого факта является появление в тексте неоценочных слов, имеющих отрицательные или положительные коннотации. Согласно энциклопедии *Кругосвет «*Коннотации являются разновидностью связанной со словом так называемой прагматической информации, поскольку отражают не сами предметы и явления действительного мира, а отношение к ним, определенный взгляд на них» (http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/KONNOTATSIYA.html). Примерами таких слов с отрицательными коннотациями в общественно-политических текстах являются слова *безработица*, *инфляция, стагнация*. В области отзывов о ресторанах слова *волос*, *майонез* несут в себе отрицательные коннотации, т.е. уже появление таких слов в текстах является признаком того, что тональность текста будет скорее отрицательной. В технической области такими словами являются слова, обозначающие поломки (*fail, crash, overload, trip, fix, mess, break*), как это указывалось в работах [12, 13].

В работах [13, 14] для автоматического выявления слов, имеющих отрицательные или положительные коннотации в общественно-политической области, используется специальный набор контекстов вида «бороться с», «предотвратить», «бороться за» и др.

Другой способ выявления слов, имеющих отрицательные или положительные коннотации, обсуждается в работе [15]. Авторы заметили, что слова, имеющие коннотации, практически не могут употребляться с оценочными словами противоположной направленности. Так, практически невозможно сказать: *хорошая безработица, прекрасная преступность* и т.п. Поэтому предлагается для выявления таких аспектных терминов вычислять разницу частот встречаемости слов с положительными или отрицательными словами. Для улучшения качества извлечения таких аспектных терминов учитывались также отрицания, союзы, расстояние от оценочного слова до слова-аспекта.

Также в [16] указывается, что есть еще одна категория неявных оценок и аспектов, которые называются авторами «ресурсная проблема». Приводится пример: *This washer uses a lot of water (Эта посудомоечная машина расходует много воды)*. Таким образом, расходование воды является здесь аспектом, а вода – ресурсным термином, чрезмерное расходование которого является отрицательным фактом.

В [17] указывается, что ресурсные термины должны извлекаться на основе употребления с квантификаторами *много-мало*, а также рядом с глаголами потребления. В работе рассматривается итеративный алгоритм, в котором в начале задаются некоторое количество известных глаголов потребления, а также несколько известных ресурсов: газ, вода, электричество, деньги, чернила, моющее средство (detergent), мыло, шампунь.

5.3. АВТОМАТИЗАЦИЯ ВЫЯВЛЕНИЯ ПРИЗНАКОВ/СВОЙСТВ ТОВАРОВ/УСЛУГ

В качестве аспектных терминов**,** чаще всего, рассматриваются существительные и группы существительного [6, 18, 19]. Длина группы существительного предполагается не больше, чем 3-4 слова. При этом указывается, что если извлекать только отдельные существительные как аспектные термины, то они часто могут быть неоднозначными, что, например, приводит к низкому согласию между экспертами [20].

Согласно [5] существует четыре основных подхода к автоматизации извлечения аспектных терминов из текстов:

– подход, основанный на частотных существительных и группах существительного;

– подход, использующий отношения между оценочными выражениями и аспектными терминами;

– подход, основанный на машинном обучении с учителем;

– подход, основанный на статистических тематических моделях.

5.3.1. Извлечение аспектных терминов на основе частотных характеристик

Для извлечения кандидатов в аспекты большое значение имеет **частотность** их упоминания в анализируемой текстовой коллекции [4, 19]. В [21] подчеркивается, что частотные признаки работают удивительно хорошо для таких простых признаков. Вместе с тем все-таки среди частотных существительных встречается достаточно много не-аспектов, например, общелитературной лексики, кроме того плохо улавливаются малочастотные аспектные термины.

В работе [22] для извлечения аспектных терминов используется известный в информационном поиске признак **tfidf** [23], который вычисляется как на уровне документов, так и на уровне абзацев. Scaffidi et al. [24] используют для извлечения аспектных терминов **сравнение частот** именных групп в коллекции отзывов с частотами этих групп в контрастной коллекции – Национальном британском корпусе.

Если в качестве аспектных терминов извлекаются не только отдельные существительные, но и группы существительного, то необходимо использовать дополнительные признаки для более точного определения длины именной группы. Чаще всего используются так называемые **контекстные признаки**, которые оценивают частоту встречаемости словосочетания с частотой контекста. Такие признаки позволяют определить границы именной группы.

Например, в [6] используется так называемая мера FLR:

$$FLR\,(a) = f(a) \cdot R(a) \quad R(a) = \sqrt{l(a) \cdot r(a)}$$

где $f(a)$ – частота аспектного термина, $l(a)$ – количество разных слов, находящихся слева от a, $r(a)$ – количество разных слов, находящихся слева от a. Далее отбираются группы существительного с данной мерой, большей, чем в среднем для словосочетаний. Таким образом, данная мера в первую очередь отбирает группы существительного, которые имеют большое разнообразие слов на своих границах, что показывает, что анализирумый термин a не является фрагментом более длинного словосочетания.

Другим критерием, направленным к этой же цели, является известный признак C-value [25], который снижает вес данного слова или словосочетания, если оно входит в частотное словосочетание большей длины. Тем самым предполагается, что это более длинное словосочетание может рассматриваться как кандидат на аспект, а текущее представляет его фрагмент. Такой признак для отбора аспектов используется в работе [26].

В работе [27] предлагается считать аспектными терминами только те именные группы, которые появляются в виде подлежащих или объектов глаголов, или в составе предложных групп.

В работе [4] алгоритм исключает из списка потенциальных аспектных терминов те из них, которые не встречаются доста-

точно часто в заданных шаблонах, обозначающих *часть-целое* (меронимию) с целевым объектом. Для этого на основе поиска в Интернет считается показатель PMI (pointwise mutual information) встречаемости предполагаемого аспектного термина с целевым объектом. Например, для цифровых камер проверяется встречаемость кандидатов в термины в образцах вида «*of camera*», «*camera has*». Кроме того, в этой работе используется иерархию WordNet для выявления названий компонентов/частей, а также словообразовательные суффиксы типа (*-iness, -ity*). Отметим, что в какой-то мере использование WordNet, фиксированных суффиксов предполагает применение алгоритма именно к техническим областям. Подобный подход (WordNet, суффиксы) представляется неприменимым к фильмам, программному обеспечению, ресторанам. В работе [28] при обзоре работ указывается, что подход [4] является затратным по времени, поскольку идет интенсивное обращение к Интернет-поиску.

Отметим, что этот набор характеристик для извлечения аспектов (за исключением проверки на отношение меронимии в работе [4]) очень похож на характеристики, используемые для извлечения терминов в заданной предметной области [29].

5.3.2. Отношения аспектов с оценочными словам. Итеративные методы для извлечения аспектных терминов

Во многих работах указывается, что аспектный термин должен входить в шаблоны с оценочными словами [18] или хотя бы употребляться в одном и том же предложении с оценочными словами [6, 18]; также могут использоваться меры, учитывающие оба эти фактора [18].

В работе [30] для извлечения отношений между аспектными терминами и оценочными словами используется синтаксический анализатор. Отношение между аспектом и оценочным словом извлекаются на основе заданных путей синтаксической зависимости. Так, например, в предложении “*This movie is not a masterpiece*“ слова *movie* и *masterpiece* будут размечены соответственно аспектом и оценочным словом, поскольку между ними существует путь в синтаксическом дереве «*NN - nsubj - VB - dobj – NN*».

Для извлечения аспектных терминов с учетом их отношений с оценочными словами часто используются итеративные методы (bootstrapping). В качестве начального множества могут использоваться частотные именные группы, которые предполагаются аспектами, либо задаются вручную.

В известной работе [19] начальное множество аспектных терминов (частотные слова и именные группы) используется для выявления ассоциативных правил – т.е. шаблонов, посредством которых аспекты обычно связаны с оценочными словами. После получения таких правил извлекаются менее частотные аспектные термины, т.е. те именные группы, которые появлялись именно в таких шаблонах с оценочными словами.

В работе [28] рассматривается подход двойного распространения (double propagation) к извлечению аспектных терминов и расширению словаря оценочных слов. В качестве исходного множества задается небольшой словарь оценочных слов, также задаются синтаксические шаблоны, в которые обычно входят оценочные слова и аспектные термины. В итоге, вхождение известного оценочного слова в такой шаблон помогает извлекать аспект, а известный аспект, входящий в такой шаблон помогает извлекать оценочное слово.

Для очистки полученного множества аспектов применяется ряд правил. Например, предполагается, что в одном фрагменте предложения без запятых содержится только один аспектный термин, а другой кандидат должен быть удален, удаляется менее частотный в коллекции.

Оценка этого метода проводилась на пяти областях; была получена средняя F-мера – 0.85, Отметим, что эксперименты проводились на небольшом числе отзывов – в среднем 62.8 отзыва из каждой области [29].

В [21] указывается, что итеративные методы, основанные на отношениях с оценочными словами, могут находить низкочастотные аспекты. Вместе с тем извлекается достаточно много неаспектов, которые подошли под заданные шаблоны. При создании комбинированных методов, сочетающих шаблоны и частотность, начинают теряться низкочастотные аспекты и возрастает число параметров для настройки.

В [8] указывается, что метод “double propagation” для одновременного извлечения аспектов и оценочных слов, основанный на

синтаксическом пути между ними хорош для коллекции среднего размера: для маленьких коллекций метод дает пониженную полноту, в то время как для больших коллекций – в заданные синтаксические шаблоны проникает много шума.

В работе [31] для оценки значимости аспектных терминов вводятся еще два фактора. Первый фактор рассматривает, насколько разнообразны оценочные слова применяемые к аспекту-кандидату – разнообразие обычно свидетельствует о значимости аспектного термина. Во-вторых, в коллекции ищется подтверждение связи аспектного термина с сущностью посредством заданных шаблонов. Например, в области автомобилей можно найти такие фразы как *the engine of the car, the car has a big engine,* которые свидетельствуют об отношении *часть-цел*ое между *engine* и *car*. Если слово одновременно встречается и с оценочным словом, и в отношениях с заданной сущностью, то это дает этому аспекту-кандидату сразу высокий вес: например, "*there is a bad hole in the mattress*".

В работе [6] для итеративного поиска аспектных терминов используется некоторое начальное множество аспектов, которое пополняется на основе:

– учета меры взаимной информации нахождения аспекта кандидата в одних и тех же предложениях, что и аспекты из начального множества аспектов и частотности аспекта-кандидата,

– при пополнении аспектов полезна очистка избыточных аспектов – например, если в множество аспектов входит и более короткий аспект.

Число вручную выделяемых аспектов у товара может достигать до 200 аспектов в технических областях. F-мера выделяемых аспектов в данной работе порядка 72.9%. Обучение проводилось на 45-100 текстов для отдельного объекта [6].

5.3.3. Использование методов машинного обучения для выявления аспектных терминов

Имеется два направления использования методов машинного обучения с учителем для выявления аспектных терминов:

– методы, основанные на предварительном составлении списка аспектных терминов в некоторой предметной области, и обучение модели, использующей перечисленные в предыдущих разделах признаки, присущие аспектам;

– методы, основанные на разметке последовательности слов в отзывах (разметка аспектных терминов, оценочных слов)

В работе [32] для извлечения аспектных терминов помимо частотности аспектов-кандидатов в отзывах используется сопоставление кандидатов с заголовками словарных статей в Википедии, и семантическая близость кандидатов, рассчитанная на основе совокупностей ссылок соответствующих статей Википедии (в итоге 2 признака), а также ассоциирование кандидата в аспекты с именем сущности при поиске в Интернете. Результат извлечения аспектов для нескольких объектов оценивается как 72.7% F-меры.

В работе [20] в качестве набора признаков для извлечения аспектных терминов в виде отдельных существительных из отзывов о лаптопах на русском языке рассматривается следующий набор признаков:

– частотность в коллекции отзывов,

– близость к оценочным словам (окно величиной p), в данном случае рассматривалась не близость к оценочным словам в коллекции, а близость на расстоянии 3 к словам *хороший/плохой* в выдаче результатов поиска Яндекса,

– признак странности, вычисляющий относительную частоту слова по сравнению с контрастной коллекцией,

– признак tfidf,

– мера взаимной информации pmi, которая учитывается совместную встречаемость между существительными кандидатами и заявленным типом товара (лаптоп).

На основе различных вариантов каждой из мер, авторами работы было получено 23 признака. Указывается, что результат извлечения близок к результатам англоязычных работ, которые заявляют о F1-мере 0.76-0.86 для разных областей.

Однако наиболее популярными в области извлечения аспектных терминов на основе методов машинного обучения являются подходы, основанные на последовательном разметке, при которой аспекты и не-аспекты аннотируются в корпусе. К размеченным данным применяются методы вида HMM (Hidden Markov models) и CRF (Conditional Random Fields) [33, 34]. В качестве признаков используются такие характеристики как собственно слова, части речи, синтаксические зависимости, расстояния, предложения с оценочными словами и др. Эти же модели могут применяться и для совместного извлечения аспектов и оценочной лексики.

В [21] указывается, что методы, основанные на машинном обучении, могу выявлять и низкочастотные аспекты, но требуют разметки данных. Конечно, особенно большие трудозатраты требуются для разметки данных для последовательных методов машинного обучения.

5.3.4. Использование тематических моделей для извлечения аспектных терминов

Извлечение аспектов может моделироваться так называемыми статистическими тематическими моделями, т.е. методами, которые предполагают, что каждый текст состоит из набора тем, а каждая тема представляет собой вероятностное распределение слов. Обычно рассматриваются два типа тематических моделей: pLSA (probalistic Latent Semantic Analysis) и LDA (Latent Dirichlet Allocation) [35, 36]. В результате применения тематических моделей к коллекции текстов порождается совокупность тем, каждая из которых представляет собой список слов с вероятностями их отнесения к этой теме.

Для извлечения аспектов необходима модификация базовых тематических моделей, направленная на то, чтобы отделить оценочные слова и топики в отдельные темы. При успешном применении таких моделей происходит два одновременных действия: извлечение аспектов и их группирование в обобщенные категории аспектов.

Одна из известных модификаций базовой модели LDA для извлечения аспектных терминов описана в работе [37], которые показали, что применение базовой модели LDA, которая строится на информации о взаимной встречаемости слов в одних и тех же текстах, не является эффективной для извлечения аспектов, поскольку во множестве разных отзывов может содержаться один и тот же набор аспектов. Они применяют глобальную модель для извлечения именований сущностей, а для извлечения аспектных терминов используют скользящее окно из слов или предложений (например, 3 предложения). Собственно, встречаемость слов в таких фрагментах используется для выявления аспектов, при этом они не различают аспектные термины и оценочные слова. В статье приводится следующий пример топика «Обслуживание»: *staff,*

friendly, helpful, service, desk, concierge, excellent, extremely, hotel, great, reception, English, pleasant, help.

В работе [38] предложена гибридная модель MaxEnt-LDA (комбинация моделей Maximum Entropy и LDA), в которой производится совместное извлечение аспектных и оценочных слов на основе синтаксических признаков, помогающих разделить аспектные и оценочные слова. Метод Maximum Entropy используется для подбора параметров на размеченных данных.

В [16] указываются следующие проблемы применения тематических моделей для извлечения и группирования аспектных терминов:

– требуются большие объемы данных и тщательная настройка параметров моделей для получения достаточно качественных результатов,

– методы основаны на семплировании Гиббса и поэтому каждый раз дают несколько иной результат,

– тематические модели легко выявляют частотные аспекты, которые выявляются и многими другими методами.

5.3.5. Группирование аспектов

Выделенные аспектные термины могут быть достаточно разнообразными, и для удобства пользователя они обычно группируются в обобщенные категории. Такими категориями для ресторана могут быть: «Кухня», «Интерьер», «Обслуживание», «Местоположение». При этом обобщенный аспект «Кухня» объединяет множество блюд и продуктов питания, которые могут предлагаться в том или ином ресторане.

В [16] указывается, что автоматизация группировки аспектов является критической для многих приложений анализа тональности отзывов.

Использование общезначимых словарей синонимов и тезаурусов имеет в этой задаче ограниченное применение, поскольку такие группировки аспектных терминов существенно зависят от предметной области. Кроме того, часто аспектные термины выражаются словосочетаниями, которые обычно не описываются в словарях.

В работах [39, 40] предложен алгоритм частичного обучения, который разбивает аспектные термины на предопределенные категории аспектов. При этом предполагается, что сами по себе

аспектные термины уже выделены каким-то методом. Сначала авторы вручную относят небольшое количество аспектных терминов к категориям. Затем применяют Expectation Maximization (EM) алгоритм для работы с размеченными и неразмеченными примерами. Кластеризация проводится на базе сходства контекстов упоминания аспектных в окне 15 слов налево и направо. Если в окне встречается другой аспектный термин, то он не включается в окно. Также исключаются стоп-слова.

В методе также применяются два вида дополнительной информации для лучшей инициализации EM-алгоритма: аспектные термины в виде именных групп, имеющие общие слова, обычно относятся к одной категории аспектов (*battery life* и *battery power*), и аспектные термины, являющиеся синонимами в словаре, также чаще всего будут принадлежать одной группе. Эти две эвристики позволяют EM-алгоритму достигать лучших результатов.

Данный алгоритм и различные другие варианты кластеризации аспектных терминов тестируются на нескольких предметных областях. Лучший результат качества кластеризации равный 0.55 по мере Purity получен в этой работе на основе EM алгоритма. Мера Purity – это мера в кластеризации, замеряющая долю максимального эталонного кластера в автоматических кластере, которая затем усредняется по всем автоматическим кластерам. Таким образом, на текущий момент лучший метод кластеризации в состоянии лишь приблизительно наполовину повторить эталонную кластеризацию.

В работе [41] ставится задача выстроить иерархическую классификацию аспектных терминов, подобно экспертной классификации. Иерархия аспектов строится на основе нескольких признаков сходства:

– контекстный признак: два слова влево и вправо,

– признак совместной встречаемости аспектных терминов, вычисляемый на основе меры взаимной информации PMI,

– длина синтаксического пути между аспектными терминами в предложении, а также синтаксические роли в предложениях (подлежащее, объект, модификатор и т.п.),

– лексические признаки, включая извлеченное из Интернета определение аспекта.

Иерархия строится итеративно, на основе минимизации нескольких критериев (minimum Hierarchy Evolution, minimum

Hierarchy Discrepancy, minimum Semantic Inconsistency), веса признаков подбираются на основе 50 иерархий WordNet и ODP (Open Directory Project).

Результаты показывают, что если начальная иерархия совсем не задана, то качество получаемой иерархии в среднем 0.3-04 F-меры. Если задано 20% иерархии, то качество составляет 0.4-0.5 F-меры. Среди признаков максимальный вклад у меры совместной встречаемости.

Ранее обсуждалось, что статистические тематические модели могут одновременно извлекать и группировать аспекты. Для учета в этих моделях знаний о предметной области в работе [42] предложено использовать дополнительные ограничения, извлекаемые из онтологии предметной области, которые могут улучшить качество создаваемых кластеров. Ограничения носят форму must-links и cannot-links. Must-links определяют, что два слова должны быть в одном кластере, cannot-links задают, что два слова не могут быть в одном кластере. Однако предложенный метод приводит к экспоненциальному росту в кодировании cannot-links и имеет сложность в обработке большого количества ограничений.

В работе [43] знание о предметной области сообщается в виде тематической модели в виде исходных (seed) слов для каждой категории аспектов. Кроме того, модель разделяет аспекты и оценочные слова. Приводятся следующие примеры исходных слов:

– *Staff (staff, service, waiter, hospitality, upkeep)*
– *Cleanliness (curtains, restroom, floor, beds, cleanliness)*
– *Comfort (comfort , mattress, furniture, couch pillows)*

Оценка подхода показывает, что два заданных слова в аспекте приводит, в среднем, к качеству извлечения аспектных слов, измеряемых мерой точности на заданном уровне 30 слов: P@30=70%, пять заданных слов – P@30=77%.

5.4. ОПРЕДЕЛЕНИЕ ТОНАЛЬНОСТИ ПО ОТНОШЕНИЮ К АСПЕКТНЫМ ТЕРМИНАМ

Как и в общей задаче анализа тональности по документам и предложениям возможно использование двух групп основных методов: методов машинного обучения и инженерно-лингвистических методов.

Ключевой вопрос при проставлении оценок тональности аспектов заключается в том, как определить диапазон действия каждого оценочного выражения, относится ли оценочное выражение к аспекту, упомянутому в этом предложении [5]. Одно из основных направлений решения этой проблемы базируется на использовании синтаксической структуры предложений в форме деревьев зависимости [3, 5, 7].

5.4.1 Методы машинного обучения для определения тональности по отношению к аспектам

В работе [7] на основе заранее собранных и вычитанных оценочных слов и аспектов задача проставления оценок аспектам рассматривается как задача классификации, т.е. для заданного предложения классификатор должен проставить, к какому именно аспектному термину относится данное оценочное слово, что может быть существенным для длинного предложения, в котором упомянуто несколько оценок и несколько аспектов (*хорошая пицца, но лазанья была ужасная*).

В качестве признаков рассматриваются следующие:

– признаки расположения: расстояние между аспектным термином и оценочным словом, число аспектов и оценочных слов в предложении, длина предложения, пунктуация, наличие одних аспектов между другими аспектами и оценочными словами, порядок расположения аспекта и оценочного слова,

– лексические признаки: набор слов между аспектным термином и оценочным словом, наличие союзов и др.,

– части речи оценочного слова и аспектного термина, набор тегов частей речи между аспектом и оценочным словом, части речи соседних слов,

– признаки, основанные на синтаксической структуре: набор тегов по пути между аспектом и оценочным словом, близость по синтаксическому дереву.

В экспериментах было показано, что все четыре типа признаков существенны для выделения пары аспектный термин – оценочное слово, достигнутая F-мера составила 82.2%. Базовый уровень для сравнения, состоявший в том, что оценочное слово приписывается к ближайшему аспекту, составил – 76.6% F-меры. Авторы подчеркивают, что они ожидали, что прирост будет больше.

В работе [3] рассматривается задача анализа тональности твитов по отношению к заданному объекту. Задача определения тональности решается с помощью двух этапов. На первом этапе на основе признаков типа эмотиконы, пунктуация, полярность слов и др. определяется, является ли предложение объективным или субъективным. А на втором этапе определяется тональность по отношению к заданному объекту.

Для второго этапа на основе обучающей выборки и проведенного синтаксического анализа порождаются признаки вида: если w_i – переходный глагол и T – его объект, то порождается признак *wi_arg2,* где *arg* – сокращение для синтаксического аргумента, например, фраза «*I love iPhone*» порождает признак *love_arg2.* Подобным же образом фиксируются признаки подлежащего при глаголе, *T* как главное для зависимого слова, *T*, соединенное связочным глаголом со словом. Также учитывается отрицание в предложении. Правильность классификации (Accuracy) твитов по заданным объектам после выполнения машинного обучения на указанных признаках достигает 85.6%, без учета синтаксической связи правильность классификации – 78.8%.

5.4.2. Лингвистико-инженерные методы проставления оценок аспектам

В лингвистико-инженерных методах предполагается, что на момент классификации известны:

– названия сущностей, их аспектов;

– имеется словарь оценочных слов и выражений, а также правила их преобразования в зависимости от контекста и правила суммирования.

Обработка идет обычно по предложениям и включает в себя несколько этапов [16].

Сначала производится проставление в предложении известных аспектных терминов и оценочных слов; оценочные слова имеют проставленную в словаре оценку тональности – в простейшем случае {1, -1}. К оценочным словам применяются операторы, которые могут менять тональность оценочного слова на противоположную.

Далее необходимо учесть структуру предложения для возможной модификации базовых оценок. В частности, в работе [45] указывается на важность обработки союзов типа *но*, *однако*. Если

во второй части предложения не обнаружено оценочных слов, но присутствуют союзы и *но* или *однако*, то второй части предложения должна быть приписана оценка, противоположная оценке первой части предложения.

В результате должно быть проведено агрегирование оценок по каждой аспектной категории. В работе [45] предлагается следующая процедура проставления оценок аспектов в отдельном предложении. Пусть в предложении s содержится набор аспектных терминов $\{a_1, ..., a_n\}$ и оценочных выражений $\{sw_1, ..., sw_n\}$, для которых оценки из словаря уже модифицированы с учетом операторов и контекста. Тогда оценки тональности каждого аспектного термина вычисляются по следующей формуле:

$$score(a_i, s) = \sum_{swj \in s} \frac{sw_j so}{dist(sw_j, a_i)}$$

где sw_j – оценочное слово или выражение, $sw_j so$ – числовая оценка тональности sw_j, $dist(sw_j a_i)$ – расстояние между оценочным словом и аспектом. Таким образом, к каждому аспектному термину в предложении приписываются все оценки, упомянутые в этом предложении, однако их вес падает в зависимости от расстояния между аспектом и оценкой. Если окончательный вес – положительный, то и оценка аспекта положительная, отрицательный вес означает отрицательную оценку, вес 0 – нейтральную оценку.

Результаты, представленные в [45], использующие вышеуказанную формулу, учет операторов, обработку союза *но* и учет контекстно-зависимых оценочных слов достигает F-меры 91% на 5 предметных областях. Система Opine на этих же данных получает 87% [4], алгоритм [20] – 83%.

В работе [45] используется шесть правил композиции оценок для определения тональности по отношению к объекту: *конверсия тональности, агрегация, распространение, доминирование, нейтрализация* и *интенсификация*.

Конверсия – это применение отрицаний и перевод в противоположную тональность. *Агрегация* применяется для синтаксических групп вида *adjective-noun, noun-noun, adverb-adjective, adverb-verb,* имеющих противоположную тональность, например, *beautiful fight*. В таком случае, этой фразе приписывается

доминирующая тональность модификатора: POS(‘*beautiful*’) & NEG(‘*fight*’) =>POSneg(‘*beautiful fight*’).

Правило распространения применяется, когда в предложении употребляется глагол распространения или передачи: PROP-POS(“*to admire*”) & “*his behavior*” => POS(“*his behavior*”); “*Mr. X*” & TRANS(“*supports*”) & NEG(“*crime business*”) => NEG(‘*Mr. X*’).

Правило доминирования заключается в том, что если полярности глагола и его объекта различны, то полярность глагола преобладает (e.g., NEG(“*to deceive*”) & POS(“*hopes*”) => NEG(“*to deceive hopes*”)); если в сложном предложении фразы соединены союзом «но», то тональность второй части предложения доминирует: например: ‘NEG(“*It was hard to climb a mountain all night long”*), *but* POS(“*a magnificent view rewarded the traveler at the morning”*).’ => POS(предложение))

Правило нейтрализации применяется, когда предлог-модификатор или оператор условия относится к тональному выражению, например, “*despite*” & NEG(‘*worries*’) => NEUT(“*despite worries*”). *Правило интенсификации* усиливает или ослабляет вес тональности, например, Pos_score(“*happy*”) < Pos_score(“*extremely happy*”)).

5.5. ТЕСТИРОВАНИЕ ОБЪЕКТНО-ОРИЕНТИРОВАННЫХ СИСТЕМ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТОВ

Задача автоматического анализа тональности текстов является сложной комплексной проблемой. Поэтому организуются различные открытые тестирования подходов к анализу тональности текстов. В состав таких тестирований входят такие как Блог трек, проводимый в рамках конференции TREC, в котором нужно по запросу найти мнение пользователя о сущности, упомянутой в запросе [46]; задания конференции TAC под названием Opinion QA Tasks [47], включающие нахождение ответов на вопросы, содержащие мнения; задания анализа мнений на конференции NTCIR, посвященной обработке текстов на восточных языках [48] и др.

В подразделах данного раздела мы рассмотрим тестирования, связанные с анализом тональности текстов по отношению к некоторому объекту:

– аспектный анализ тональности отзывов, проводимый в рамках конференций SemEval 2014–2015 [49],

– задания анализа Твиттера, связанные с проблемой управления репутацией [2, 50],

– а также организованное в 2014–2015 годах тестирование систем объектно-ориентированного анализа тональности для текстов на русском языке SentiRuEval [51].

5.5.1. Анализ тональности отзывов в рамках конференции SemEval

Впервые оценка систем тональности отзывов по отношению к заданным аспектам сущности была организована в рамках конференции SemEval в 2014 году [52]. Данные для обучения и тестирования включали изолированные предложения, извлеченные из отзывов, а не полные отзывы в двух предметных областях: ресторанах и лаптопах. 3K предложений было подготовлено для обучения в каждой из областей. Множество аспектных категорий по ресторанам включали: *food (еда), service (обслуживание), price (цена), ambience (обстановка, атмосфера), anecdotes/miscellaneous (другое)***.**

В 2015 тестирование обработки отзывов в рамках SemEval (http://alt.qcri.org/semeval2015/task12/) включает уже полные отзывы. Аспектные категории усложняются и теперь уже состоят из пар сущность-характеристика (Entity-Atribute pairs – E#A). Набор пар E#A включает в области ресторанов шесть типов сущностей (RESTAURANT, FOOD, DRINKS, SERVICE, AMBIENCE, LOCATION) и 5 типов атрибутов (GENERAL, PRICES, QUALITY, STYLE_OPTIONS, MISCELLANEOUS). Область лаптопов содержит 22 типа сущностей and 9 типов атрибутов (GENERAL, PRICE, QUALITY, OPERATION_PERFORMANCE и др.). Примеры аннотирования предложений в области отзывов о ресторанах выглядят следующим образом:

1) Great for a romantic evening, but over-priced. → {AMBIENCE#GENERAL}, {RESTAURANT#PRICES}

(2) The fajitas were delicious, but expensive. → {FOOD#QUALITY}, {FOOD# PRICES}

5.2. Тестирование систем анализа репутации

В 2012–2014 годах в рамках конференции CLEF был организован семинар RepLab, посвященный тестированию систем анализа репутации [2, 51]. Суть исследования заключалась в отслежива-

нии положительных и отрицательных сообщений в Твиттере для заданной сущности (компании или персоне).

Для участников мониторинг включал два задания:

– разрешение неоднозначности. Нужно было извлечь из потока сообщений Твиттера те твиты, которые относятся к заданным объектам (компаниям, персонам). Тестирование основывалось на ручной разметке (релевантный, нерелевантный, неопределенный);

– полярность сообщения для заданного объекта. Системы должны были извлечь последствия для объекта, которые следуют из данного сообщения. Ручная оценка включала теги: положительный/отрицательный/нейтральный/неопределенный.

Подчеркивается, что для анализа репутации системы не должны классифицировать сообщения на факты и мнения. Цель систем – найти полярность для репутации, независимо от того, является ли высказывание фактом или мнением.

За два года тестирования участниками был получен значительный рост качества результатов классификации твитов по полярности по отношению к заданному объекту: в 2012 году наилучший достигнутый показатель правильности классификации (accuracy) составлял 0.41, в 2013 году – 0.69. Организаторы [51] объясняют этот прогресс тем, что за два года тестирования был собран и размечен огромный корпус твитов (более 140 тысяч) на двух языках, что позволило с большей надежностью применять стандартные методы машинного обучения, в то время как сами поставленные задачи остаются достаточно проблематичными для современных систем автоматической обработки текстов.

5.5.3. Тестирование SentiRuEval

Мероприятие по оценке систем анализа тональности для текстов на русском языке SentiRuEval, которое организовано в 2014–2015 гг. является вторым после сравнительных исследований систем анализа тональности в рамках семинара по информационному поиску РОМИП, организованному в 2011–2013 годах. Тестирование в рамках РОМИП было направлено на выявление общей тональности текста (отзыва, поста в блоге, новостной цитаты) [53, 54]. Новое тестирование SentiRuEval направлено на исследование методов анализа текстов по отношению к некоторому заданному объекту или его характеристикам [51].

В 2014–2015 годах в рамках SentiRuEval имеется два типа задания: объектно-ориентированный анализ твитов для двух типов организаций (банки и телекоммуникационные компании) и аспектно-ориентированный анализ отзывов пользователей в двух предметных областях (рестораны и автомобили).

5.5.3.1. Тестирование анализа тональности отзывов по аспектам в рамках SentiRueval

Каждый отзыв содержит мнения пользователя о конкретном объекте. Такие мнения структурируются по заранее заданному набору *целевых аспектов*, т. е. составных частей, либо характеристик оцениваемого объекта. Для ресторанной тематики такими аспектами являются: *кухня*, *интерьер, сервис, цена*. Для автомобилей список аспектов включает в себя: *безопасность*, *комфорт*, *надежность*, *внешний вид*, *цены, ходовые качества*. Набор целевых аспектов дополнен аспектом «*объект в целом*», представляющим общее мнение об объекте.

Для создания обучающей коллекции была осуществлена разметка отзывов, при которой в тексты вносилась следующая информация:

– выделяются аспектные термины, включая эксплицитные, имплицитные и тональные факты;

– выделенным аспектным терминам приписывается их тональность: позитивный, негативный, противоречивый (both) и нейтральный;

– выделенные аспектные термины относятся к аспектной категории;

– отмечается статус выделенного аспектного термина относительно текущего мнения: релевантный (REL), относится к прошлому мнению автора или других людей (PREV), относится к другому объекту (CMPR), относится к гипотетической ситуации (IRR), ирония (IRN). Такая разметка помогает выявить аспектные термины, учет которых скорее затрудняет качество анализа, поскольку они не относятся к текущему мнению автора;

– приписывается оценка аспектной категории в целом по отзыву: нейтральный, положительный, отрицательный, противоречивый, оценка отсутствует.

Участники на выбор могли решать указанные ниже задачи.

Задача А. Выделение **релевантных отзыву** эксплицитных аспектных терминов. При этом не должны размечаться как эксплицитные аспектные термины упоминания аспектов, относящихся к другим объектам или ситуациям, упоминаемым в отзывах.

Задача Б. Выделение **релевантных отзыву** всех аспектных терминов, включая неявные аспектные термины и тональные факты.

Задача В. Присваивание оценки тональности **эксплицитным** аспектным терминам.

Задача Г. Присвоение аспектной категории **эксплицитным** аспектным терминам.

Задача Д. Заполнение оценок аспектных категорий по отзывам в целом.

Для каждой задачи организаторами были подготовлены прогоны, представляющие базовые уровни (baseline) для сравнения, т.е представляющие собой очень простые решения поставленных задач.

Базовая система для задач А и Б извлекает список размеченных терминов из обучающей коллекции, лемматизирует их и размечает их в тестовой коллекции на основе ее лемматизированного представления. Если к некоторой последовательности слов применимо более одного термина, то предпочитается более длинный термин.

Базовая система задачи В приписывает аспектному термину его наиболее частотную аспектную категорию, на основе информации из обучающей коллекции. Если термин отсутствует в обучающей коллекции, то приписывается наиболее частотная аспектная категория. Базовая система задачи Г приписывает аспектным терминам тональности на основе таких же принципов. Базовый уровень для задачи Е представляет собой наиболее частую категорию тональности для каждой аспектной категории (во всех случаях это была положительная тональность).

В тестировании приняли участие 11 участников, причем задача анализа отзывов о ресторанах привлекла значительно больше внимания, чем отзывы об автомобилях. Как указывается в [51], лучшие результаты, полученные участниками для задач А и Б, пока не намного превзошли базовый метод извлечения аспектных терминов, переносящий разметку из обучающего множества в тестовое. Например, при точном сопоставлении эксплицит-

ных аспектов по ресторанам лучший результат составил – 0.632 F-меры, а baseline результат – 0.608. Многие участники не смогли превзойти результат baseline системы.

Задачи В и Г являются задачами классификации аспектных терминов, и лучшие результаты были получены на основе методов машинного обучения SVM и Gradient Boosting. Обучающие и тестовые данные, результаты участников, а также скрипты для подсчета результатов доступны по адресу: http://goo.gl/Wqsqit.

5.5.3.2 Тестирование систем анализа твитов относительно заданного объекта

В коллекцию твитов для автоматического анализа тональности собраны твиты, упоминающие названия крупнейших банков и крупнейших телекоммуникационных компаний. В этом тестировании специально собраны организации двух важнейших для анализа тональности сфер деятельности, чтобы изучить вопросы зависимости анализа тональности твитов от конкретной предметной области.

Наборы твитов были подготовлены в форме xml-файлов, в которых после текста твита перечислены все организации, по которым извлекались твиты, и та организация, которая была обнаружена именно в данном конкретном твите, помечена значением *0*:

<column name=”sberbank”>0</column>

что означает, что по умолчанию установлено значение тональности твита – нейтральное. В задачу систем входило следующее: в зависимости от содержания твита для тех организаций, которые были отмечены значением 0, изменить тональность на положительную или отрицательную, или оставить без изменений. Таким образом, в SentiRuEval в отличие от ранее описанного тестирования RepLab (см. п. 5.2.) не было задачи фильтрации твитов на относящийся к заданной организации и нерелевантные.

Разметка твитов производилась по пяти категориям, а именно разметчики должны были проставлять следующие отметки выделенным сущностям:

- 1, если тональность – положительная,
- -1, если тональность – отрицательная,
- оставить 0, если тональность нейтральная,
- -- (два минуса), если нечитабельно, непонятно, не относится к названным темам и т.п.,

• +- (плюс-минус), если в твите выражены обе оценки; или твит настолько ироничен, что тональность его неясна.

Два последних типа твитов были далее исключены из тестирования, из-за их нерелеватности заданным сущностям или непонятности тональности для человека.

В качестве обучающей и тестовой коллекции было размечено по 5 тысяч твитов для банков и телекоммуникационных компаний (т.е. всего 20 тысяч твитов)

Анализ разметки обучающей коллекции показал, что при разметке ряда твитов возникает серьезное расхождение мнений по поводу влияния на репутацию или нейтральности некоторых твитов. Например, при анализе твита «*я сегодня ходил в сбербанк за картой, там оч милая девушка работала*» мнения обсуждающих разделились пополам: часть людей считала, что твит нейтральный по отношению к Сбербанку, а часть считала, что этот твит позитивный.

Для того чтобы снизить проблему субъективности разметки, а также фактор случайной ошибки, то тестовая выборка размечалась тремя разметчиками и результирующая оценка тональности проставлялась методом голосования.

Отметим, что временные отрезки для обучающей и тестовой выборки твитов различались. Твиты обучающей выборки извлекались осенью 2014 года, а твиты для тестовой выборки относятся к 2013 году.

Лучшие результаты классификации твитов оказались не очень высокими и составили 0.488 макро F-меры для телекоммуникационных компаний, и 0.360 макро F-меры для банков. Интересно, что лучшими подходами по классификации твитов для компаний оказались представители двух разных типов подходов – машинного обучения (SVM), и подхода, основанного на правилах, без всякого машинного обучения. Оба лучших подхода использовали информацию о синтаксических отношениях между словами в твите.

Дополнительно, один из участников выполнил независимую экспертную разметку высланных тестовых твитов и получил результаты 0.703 Макро F-меры, что может рассматриваться как максимум качества, которого могут достигнуть системы. Видно, что участникам еще много нужно работать для того, чтобы приблизиться к результатам человека. Обучающая и тестовая кол-

лекции, результаты участников и скрипты для подсчета метрик доступны по адресу http://goo.gl/qHeAVo.

Заключение

В течение последних 10–15 лет задача автоматического анализа тональности текстов вызывает неизменно высокий интерес у исследователей и имеет разнообразные сферы применения на практике. В данной статье были рассмотрены подходы к задачам, связанным с анализом тональности по отношению к заданному объекту, а также к его характеристикам. Также мы описали открытые тестирования, проводившиеся в этой сфере для систем анализ тональности текстов на английском и русском языках. Обучающая и тестовая коллекции, результаты участников и скрипты для подсчета метрик опубликованы для некоммерческого использования.

Благодарности

This work is partially supported by RFBR grant 14-07-00682.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Pang B., Lee L., Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques // *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*. – Vol. 10. – 2002. – P. 79–86.

2. Amigo E., Corujo A., Gonzalo J., Meij E., Rijke M. Overview of RepLab 2012: Evaluating Online Reputation Management Systems // *International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF-2012)*. – Berlin: Springer Berlin Heidelberg, 2012. – P. 333–352.

3. Long J., Mo Yu, Ming Zhou, Xiaohua Liu, Tiejun Zhao. Target dependent twitter sentiment classification // *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*. – 2011. – P. 151–160.

4. Popescu, A., Etzioni O. Extracting product features and opinions from reviews // *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2005)*. – 2005.

5. Liu B., Zhang L. A survey of opinion mining and sentiment analysis // *Mining Text Data*. – Springer US, 2012. – P. 415–463.

6. Bagheri A., Saraee M., de Jong F. An Unsupervised Aspect Detection Model for Sentiment Analysis of Reviews // *Natural Language Processing*

and Information Systems. – Berlin: Springer Berlin Heidelberg, 2013. – P. 140–151.

7. Glavaš G., Korencic D., Šnajder J. Aspect-Oriented Opinion Mining from User Reviews in Croatian // *Proceedings BSNLP workshop, ACL 2013*. – 2013.

8. Zhang L., Liu B. Aspect and Entity Extraction for Opinion Mining // *Data Mining and Knowledge Discovery for Big Data*. – Berlin: Springer Berlin Heidelberg, 2014. – P. 1–40.

9. Liu B. Sentiment analysis and Subjectivity // *Handbook of Natural Language Processing*. – CRC Press, Taylor and Francis Group, Boca Raton, 2010. – P. 1–38.

10. Gupta N. K. Extracting phrases describing problems with products and services from twitter messages // *Computación y Sistemas*. – Vol. 17, Numb. 2. – 2013. – P. 197–206.

11. Ivanov V., Tutubalina E. Clause-based approach to extracting problem phrases from user reviews of products // *Analysis of Images, Social Networks and Texts*. – Springer International Publishing, 2014. – P. 229–236.

12. Tutubalina E., Ivanov V. Unsupervised Approach to Extracting Problem Phrases from User Reviews of Products // *Proceedings of the Aha! workshop on Information Discovery in Texts, Coling-2014*. – 2014. – P. 48–53.

13. Feng S., Bose R., Choi Y. Learning general connotation of words using graph-based algorithms // *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. – Association for Computational Linguistics, 2011. – P. 1092–1103.

14. Feng S., Kang J. S., Kuznetsova P., Choi Y. Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning // *Proceedings of ACL*. – 2013. – P. 1774–1784.

15. Zhang Lei, Bing Liu. Identifying noun product features that imply opinions // *Proceedings of the Annual Meeting of the Association for Computational Linguistics (short paper) (ACL-2011)*. – 2011.

16. Liu B., Zhang L. A survey of opinion mining and sentiment analysis // *Mining Text Data*. – Springer US, 2012. – P. 415–463.

17. Zhang Lei, Liu B. Extracting Resource Terms for Sentiment Analysis // *Proceedings of IJCNLP-2011*. – 2011.

18. Blair-Goldensohn S., Hannan K., McDonald R., Neylon T., Reis G. A., Reynar J. Building a sentiment summarizer for local service reviews // *Proceedings of WWW Workshop on NLP in the Information Explosion Era*. – 2008.

19. Hu M., Liu B. Mining and summarizing customer reviews // *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*. – ACM, 2004. – P. 168–177.

20. Марчук А. А., Уланов А. В., Макеев И. В., Чугреев, А. А. Автоматическое извлечение параметров продуктов из текстов отзывов при помощи интернет-статистик // *Труды Международной конференции Компьютерная лингвистика и информационные технологии Диалог-2013*. – Т.2. – 2013. – С. 81–91.

21. Moghaddam S., Ester M. Aspect-based Opinion Mining from Online Reviews // *Tutorial at SIGIR-2012*. – Portland, Oregon, USA: 2012.

22. Ku Lun-Wei, Yu-Ting Liang, Hsin-Hsi Chen. Opinion extraction, summarization and tracking in news and blog corpora // *Proceedings of AAAI-CAAW'06*. – 2006.

23. Manning C. D., Raghavan P., Schütze H. Introduction to information retrieval. – Cambridge: Cambridge University Press, 2008.

24. Scaffidi Ch., Bierhoff K., Chang E., Felker M., Ng H., Jin Ch.. Red Opal: product-feature scoring from reviews // *Proceedings of Twelfth ACM Conference on Electronic Commerce (EC-2007)*. – 2007.

25. Frantzi K., Ananiadou S., Mima H. Automatic recognition of multiword terms: the C-value/NC-value method // *International Journal on Digital Libraries*. – Vol. 3. – Numb. 2. – 2000. – P. 115–130.

26. Zhu J., Wang H., Tsou B., Zhu M. Multiaspect opinion polling from textual reviews // *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2009)*. – 2009.

27. Hai Z., Chang K., Cong G. One seed to find them all: mining opinion features via association // *Proceedings of the 21st ACM international conference on Information and knowledge management*. – ACM, 2012. – P. 255–264.

28. Qiu G., Liu B, Bu J, Chen C. Opinion word expansion and target extraction through double propagation // *Computational Linguistics*. – 2011.

29. Loukachevitch N., Nokel M. An Experimental Study of Term Extraction for Real Information-Retrieval Thesauri // *Proceedings of Terminology and Artificial Intelligence Conference TIA-2013*. – 2013.

30. Zhuang L., Jing F., Zhu X. Movie review mining and summarization // *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2006)*. – 2006.

31. Zhang L., Liu B., Lim S., O'Brien-Strain E. Extracting and ranking product features in opinion documents // *Proceedings of International Conference on Computational Linguistics (COLING-2010)*. – 2010.

32. Kovelamudi S., Ramalingam S., Sood A., Varma V. Domain Independent Model for Product Attribute Extraction from User Reviews using Wikipedia // *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-2010)*. – 2011.

33. Niklas J., Gurevych I.. Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields // *Proceedings*

of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010). – 2010.

34. Choi Y., Cardie C. Hierarchical sequential learning for extracting opinions and their attributes // *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2010)*. – 2010.

35. Blei D., Ng A., Jordan M. Latent dirichlet allocation // *The Journal of Machine Learning Research*. – Vol. 3. – 2003. – P. 993–1022.

36. Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. – Т. 1. – № 6. – 2013. – С. 657–686.

37. Titov I., McDonald R. A joint model of text and aspect ratings for sentiment summarization // *Proceedings of ACL-08*. – Columbus, Ohio, USA: HLT, 2008. – P. 308–316.

38. Zhao Wayne Xin, Jing Jiang, Hongfei Yan, Xiaoming Li. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid // *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010)*. – MIT, Massachusetts, USA: ACL, 2010. – P. 56–65.

39. Zhai Z., Liu B., Xu H., Jia P. Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints // *Proceedings of Coling-2010*. – ACL, 2010. – P. 1272–1280.

40. Zhai Z., Liu B., Xu H., Jia P. Clustering product features for opinion mining // *Proceedings of the fourth ACM international conference on Web search and data mining*. – ACM, 2011. – P. 347–354.

41. Yu J., Zha Z. J., Wang M., Wang K., Chua T. S. Domain-assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews // *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. – Association for Computational Linguistics, 2011. – P. 140–150.

42. Andrzejewski D., Zhu X., Craven M. Incorporating domain knowledge into topic modeling via Dirichlet forest priors // *Proceedings of ICML*. – 2009.

43. Mukherjee A., Liu B. Aspect Extraction through Semi-Supervised Modeling // *Proceedings of 50th Annual Meeting of Association for Computational Linguistics (ACL-2012)*. – 2012.

44. Ding X., Liu B., Yu Ph. A holistic lexicon-based approach to opinion mining // *Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008)*. – Palo Alto, California, USA: ACM, 2008.

45. Neviarouskaya A., Prendinger H., Ishizuka M. Recognition of affect, judgment, and appreciation in text // *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010)*. – ACL, 2010. – P. 806–814.

46. Macdonald C., Santos R. L., Ounis I., Soboroff I. Blog track research at TREC // SIGIR Forum. – Vol. 44. – Numb. 1. – 2010. – P. 58–75.

47. Dang H. T., Owczarzak K. Overview of the tac 2008 opinion question answering and summarization tasks // *Proceedings of the First Text Analysis Conference*. – 2008. – P. 1–16.

48. Seki Y. et al. Overview of multilingual opinion analysis task at NTCIR-7 // *Proceedings of the Seventh NTCIR Workshop*. – 2008.

49. Rosenthal S., Nakov P., Ritter A., Stoyanov V. SemEval-2014 Task 9: Sentiment Analysis in Twitter // *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval*. – 2014.

50. Amigo E., Albornoz J.C., Chugur I., Corujo A., Gonzalo J., Martin T., Meij E., Rijke M, Spina D. Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems // *CLEF-2013 , LNCS 8138*. – Berlin: Springer Berlin Heidelberg, 2013. – P. 333–352.

51. Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E. SentiRuEval: Testing Object-oriented Sentiment Analysis Systems in Russian // *Proceedings of International Conference of Computational Linguistics and Intellectual Technologies Dialog-2015*. – 2015.

52. Pontiki M., Galanis D., Pavlopoulos J., Papageorgiou H., Androutsopoulos I., Manandhar S. SemEval-2014 Task 4: Aspect Based Sentiment Analysis // *Proceedings of International Workshop on Semantic Evaluations SemEval-2014*. – 2014. – P. 27–35.

53. Chetviorkin I. I., Braslavski P. I., Loukachevitch N. V. Sentiment analysis track at ROMIP 2011 // *Proceedings of International Conference Dialog-2012*. – 2012. – P. 739–746.

54. Chetvirokin I., Loukachevitch N. Sentiment Analysis Track at ROMIP 2012. // *Proceedings of International Conference Dialog-2013*. –Vol. 2. – 2013. – P. 40–50.

Глава 6. ЗАМЕТКИ О ТЮРКСКИХ СВЕРШЕНИЯХ*

С. Г. Татевосов

6.1. ПРЕДИКАТЫ КЛАССА СВЕРШЕНИЙ: ТЕОРЕТИЧЕСКИЕ АЛЬТЕРНАТИВЫ

Публикация основополагающей книги Д. Даути «Значение слова и грамматика Монтегю» (Dowty 1979) привела к возникновению целого ряда гипотез о внутренней структуре и интерпретации событийных предикатов класса свершений типа «открыть дверь» или «разбить окно». Современные теории, описывающие такие предикаты, различаются по целому ряду параметров, часть которых перечислена в (1):

(1) Параметры варьирования вариантов анализа глаголов класса свершений:

a. количество и состав (под)событийных компонентов,

b. тип отношения, связывающего событийные компоненты,

c. спецификация дескриптивных свойств у каждого из компонентов.

Несколько вариантов теории свершений показаны в Таблице 1.

Таблица 1

Теории событийных предикатов класса свершений

	Элементы структуры	Связь между элементами	Ограничения на структуры элементов
Dowty 1979	Деятельность + Достижение (изменение состояния)	Каузальная	Нет

* Исследование поддержано грантом РФФИ 13-06-00884а.

Продолжение таблицы

	Элементы структуры	Связь между элементами	Ограничения на структуры элементов
Rappaport Hovav & Levin 1998 и др. работы	Деятельность + Достижение (изменение состояния)	Каузальная	Дескриптивные свойства деятельности не определены
Kratzer 2000, 2005	Деятельность + Результирующее состояние	Каузальная	Нет
Rothstein 2004	Деятельность + Изменение состояния	Инкрементальная	Подсобытие изменения состояния частично упорядочено инкрементальной цепочкой
Ramchand 2008	Деятельность + Процесс + Результирующее состояние	Каузальная	Нет

Задача настоящей работы – рассмотреть *некульминирующие прочтения* предикатов класса свершений в тюркских языках и показать, что этот материал подкрепляет следующие обобщения:

(i) подсобытие деятельности должно присутствовать в семантическом представлении отдельно от подсобытия изменения состояния;

(ii) различные глаголы-свершения по-разному ограничивают отношение между подсобытиями;

(iii) свершения различаются с точки зрения внутренней упорядоченности подсобытия деятельности.

Последующее изложение организовано следующим образом. В разделе 2 обсуждается материал четырех тюркских языков: татарского (мишарский диалект), чувашского, алтайского (тубаларский, или туба диалект) и карачаево-балкарского языков. Мы показываем, что в отсутствии кульминации для предикатов класса свершений допускается два прочтения – ‘неудавшаяся попытка’ и ‘частичный успех’. Некоторые предикаты предпочитают неудав-

шуюся попытку, некоторые частичный успех, некоторые другие вовсе не допускают отсутствия кульминации. В разделе 3 предлагается общий взгляд на некульминирование и обосновывается декомпозициональный подход к анализу событийной структуры свершений. В разделе 4 выделяются два типа отношений между подсобытиями, а различие между прочтениями типа 'неудавшаяся попытка' и 'частичный успех' сводятся к свойствам этих отношений. В разделе 5 мы выдвигаем предположение, что возможность некульминирующего прочтения определяется нетривиальной семантической характеристикой: если подсобытие деятельности в структуре предиката частично упорядочено временным предшествованием и/или каузальной зависимостью, некульминирущие прочтения невозможны.

6.2. НЕКУЛЬМИНИРУЮЩИЕ СВЕРШЕНИЯ В ТЮРКСКИХ ЯЗЫКАХ

Рассмотрим примеры (1) - (6). Они позволяют сделать два наблюдения. Во-первых, только часть предикатов-свершений допускают как кульминирующее, так и некульминирующее прочтение. В примерах (2) - (5) события достигают кульминации, после которой человек рассержен, (2a), пуговица оторвана, (2b), поле вспахано, (2c) и т.д. Эти предложения, как показывает обычный тест на сочетаемость с обстоятельствами длительности типа 'за две минуты', предельны. В примерах (6) - (9), напротив, кульминация не достигается, а предложения непредельны: они лишь сообщают, что деятельность агенса протекала некоторое время и прекратилась до кульминации. Что особенно важно, свершения в примерах (b) и (c) допускают некульминирующее прочтение, тогда как в примерах (a) оно невозможно.

Кроме того, некульминирующие прочтения в (b) и в (c) различаются: в (b) речь идёт о *неудавшейся попытке* (НП), а в (c) – о *частичном успехе удавшейся* (ЧУ). НП предполагает, что агенс производит определенную деятельность, целью которой является изменение состояния пациенса, однако она прекращается до того, как изменение произошло. Попытки агенса оторвать пуговицу, разбудить человека оказываются безуспешными, и пациенс сохраняет исходное состояние. В случае ЧУ событие также не до-

стигает кульминации, но радикально отличным образом: из примеров (4c) - (6c) следует, что объект претерпевает определенное изменение.

(1) Мишарский диалект татарского языка

a. marat kErIm-nE IkE mInut ECEndA aCulander-de.
M. К-ACC два минута за сердить-PST
Марат рассердил Керима за две минуты.

b. marat sAdAf-E-n IkE mInut ECEndA Ez-dE.
M. пуговица-3-ACC два минута за отрывать-PST
Марат оторвал пуговицу за две минуты.

c. daut ker-ne ike sAgAt ECEndA sukal-a-de.
Даут поле-ACC два час за пахать-ST-PST
Даут вспахал поле.

(1) Чувашский язык

a. vaCa eki Cekunt xuSAncE CutA sUnter-c-E.
В. два секунда за свет выключать-PFV-PST
Вася выключил свет за две секунды.

b. vaCa eki minut xuSAncE petuk-na vArat-r-E.
В. два минута за П.-ACC будить-PFV-PST
Вася разбудил Петю за две минуты.

c. vaCa CirEm minut xuSAncE samalot matell-a puCtar-c-E.
В. двадцать минута за самолет модель-ACC собирать-PFV-PST
Вася собрал модель самолета за двадцать минут.

(1) Тубаларский диалект алтайского языка

a. vasʲa eki Cas-xa petJa-ny kOm-di.
В. два час-DAT Петя.-ACC хоронить-PST.3SG
Вася похоронил Петю за два часа.

b. vasʲa on minut-xa eSik-ni aC-ty.
В. десять минута-DAT дверь-ACC открывать-PST.3SG
Вася открыл дверь за десять минут.

c. vasʲa suù-ny eki minut-xa soùt-ty.
В. вода-ACC два минута-DAT охлаждать-PST.3SG
Вася охладил воду за две минуты.

(1) Карачаево-балкарский язык
a. alim eki sekunt-xa kerim-ni attyr-dy.
А. два секунда-DAT К.-ACC застреливать-PST.3SG
Алим застрелил Керима за две секунды.

b. kerim on sekunt -xa xaly-ny zyrt-ty.
К. десять секунда-DAT нить-ACC рвать-PST.3SG
Керим порвал нить за десять секунд.

c. išci eki kün-ge üj-nü oj-dy.
Рабочий два-DAT день-DAT дом-ACC сносить-PST.3SG
Рабочий снес дом за два дня.

(1) Мишарский диалект татарского языка
a. *marat kErIm-nE IkE mInut bujena aCulander-de.
М. К-ACC два минута в.течение сердить-PST
Марат пытался рассердить Керима две минуты.

b. marat sAdAf-E-n IkE mInut bujena Ez-dE.
М. пуговица-3-ACC два минута в.течение отрывать-PST
Марат позанимался отрыванием пуговицы две минуты.

c. daut ker-ne ike sAgAt bujena sukal-a-de.
Даут поле-ACC два час в.течение пахать-ST-PST
Даут попахал поле два часа.

(1) Чувашский язык
a. *vaCa eki minut CutA sUnter-c-E.
В. два минута свет выключать-PFV-PST
Вася пытался выключить свет две минуты.

b. vaCa CirEm minut petuk-na vArat-r-E.
В. двадцать минута П.-ACC будить-PFV-PST
Вася пытался разбудить Петю двадцать минут.

c. vaCa CirEm minut samalot matell-a puCtar-c-E.
В. двадцать минута самолет модель-ACC собирать-PFV-PST
Вася занимался собиранием модели самолета двадцать минут.

(1) Тубаларский диалект алтайского языка
a. *vasJa eki Cas petJa-ny kOm-di.
В. два час П.-ACC хоронить-PST.3SG
Вася занимался похоронами Пети два часа.

b. vasJa on minut eSik-ni aC-ty.
В. десять минута дверь-ACC открывать-PST.3SG
Вася пытался открыть дверь десять минут.

c. vasja eki minut suù-ny soùt-ty.
В. два минута вода-ACC охлаждать-PST.3SG
Вася поохлаждал воду две минуты.

(1) Карачаево-балкарский язык
a. *alim eki saRat kerim-ni attyr-dy.
А. два час К.-ACC застреливать-PST.3SG
Алим пытался застрелить Керима два часа.

b. kerim on minut xaly-ny zyrt-ty.
К. десять минута нить-ACC рвать-PST.3SG
Керим пытался разорвать нить десять минут.

c. išci eki kün üj-nü oj-dy.
рабочий два день дом-ACC сносить-PST.3SG
Рабочий позанимался сносом дома два дня.

К вопросу о том, почему невозможны некульминирующие свершения в примерах (a) (‘рассердить’, ‘выключить’, ‘застрелить’, ‘похоронить’), мы обратися в разделе 5. В следующих двух разделах нам предстоит сосредоточиться на некульминирующих предикатах типа ‘отрывать пуговицу’, ‘будить человека’, ‘открывать дверь’ в (6b) - (9b) и ‘пахать поле’, ‘охлаждать воду’, ‘собирать модель’, ‘сносить дом’ в (6c) - (9c). Эти предикаты ставят перед нами два вопроса. Во-первых, как возникают некульминирующие прочтения (b) - (c) в (6) - (9) и как они соотносятся с предельным прочтением в (2) - (5)? Во-вторых, чем обусловлена разница между НП-свершениями типа ‘будить человека’, ‘открывать дверь’, ‘рвать нить’ в примерах (b) и ЧУ-свершениями типа ‘пахать поле’, ‘охлаждать воду’, ‘собирать модель’, ‘сносить дом’ в примерах (c)?

6.3. КУЛЬМИНАЦИЯ И ЕЕ НЕДОСТИЖЕНИЕ

Все некульминирующие предикаты в (b) - (c) в (4) - (6) образуют естественный класс: они сочетаются с обстоятельствами длительности типа ‘два часа’, а значит, являются непредельными. Более того, они предъявляют нам тот же имперфективный пара-

докс, что и прогрессивы (Dowty 1979 и обширная последующая литература): в предложениях (b) - (c) в (6) - (9) истинность пропозиции в актуальном мире не влечет за собой истинности соответствующей пропозиции из (2) - (5). Это наблюдение подводит нас к так называемой партитивной теорией некульминирования: при некульминирующим прочтении только часть ситуации из исходного экстенсионала предиката имеет место в акуальном мире. Ситуация целиком, включая кульминацию, существует только в инерционных мирах (Dowty 1979), в мирах, образующих ветвь развития события (Landman 1992) или в любых других мирах, наилучшим образом приспособленных для разрешения имперфективного парадокса. Излагаемая ниже аргументация, как кажется, не зависит от выбора одной из альтернатив. Различные варианты партитивной теории см. в Koenig and Muansuwan 2001, Bar-el et al. 2005 и Tatevosov, Ivanov 2009

Если рассматривать глагольную группу (*v*P), лишенную показателей времени и вида, как событийный предикат, *v*P в (3b) получает (нео-дэвидсоновское) семантическое представление в (10), а его некульминирующий вариант из (7b) – анализ в (11). В (11) оператор CM (от continuation modality) можно понимать как предложенный Ф. Лэндмэном (Landman 1992) оператор прогрессива PROG, который выделяет из события его стадии, за исключением одного отличия. Поскольку одной из стадий события *e* является само *e*, PROG позволяет *e* кульминировать в актуальном мире. CM, в отличие от PROG, извлекает из события его *собственные неконечные стадии*.

(10) $\lambda e[wake(e) \wedge agent(basil)(e) \wedge theme(peter)(e)]$

(11) $\lambda e[CM(e, \lambda e'.wake(e') \wedge agent(basil)(e') \wedge theme(peter)(e')]$

(10) - (11) предлагают недекомпозициональный анализ предиката ‘будить’: событие пробуждения рассматривается как неразделимое целое, а оператор CM выделяет его собственные неконечные стадии. В этом месте возникает первая проблема.

Чтобы осознать ее природу, присмотримся внимательнее к двум типам некульминирующих свершений – НП и ЧУ. НП-предложения в (6b) - (9b) описывают событие, в котором с пациенсом ничего не происходит. ЧУ-предложения в (4c) - (6c) предполагают, что объект претерпевает некоторое изменение. Другими словами,

ЧУ отличаются от НП тем, что изменение объекта протекает в актуальном мире. Общим для ЧУ и НП является отсутствие в актуальном мире кульминации. Это показано в Таблице 2.

Таблица 2

Кульминирующие и некульминирующие прочтения

	Кульминирующие	**Некульминирующие**	
		Частичный успех	**Неудавшаяся попытка**
Деятельность агенса	В актуальном мире	В актуальном мире	В актуальном мире
Изменения в пациенсе	В актуальном мире	В актуальном мире	Не в актуальном мире
Кульминация	В актуальном мире	Не в актуальном мире	Не в актуальном мире

Интуитивно кажется, что неудавшиеся попытки в случае типа ‘будить человека’ в (7b) отличаются от частично успешных ситуаций типа ‘собирать модель’ в (7c) тем, как деятельность агенса соотносится с вызванным ею изменением состояния. ЧУ-свершения устроены так, что каждая контекстно-релевантная часть деятельности агенса вызывает изменение в пациенсе. Таким образом, в любой собственной неконечной стадии ситуации в целом изменения в пациенсе неизбежны. В случае НП-свершений, напротив, неконечные элементы деятельности не влекут изменений. Если бы деятельность ‘будить’ кульминировала, всё изменение состояния наступило бы в конечной ее части. Деятельность, однако, она прекращается до кульминации, когда пациенс ещё находится в исходном состоянии.

Проблема заключается в том, что это интуитивное описание невозможно сделать эксплицитным при недекомпозиционном анализе предикатов класса свершений, как в (10) - (11). Предикаты, описывающие события собирания модели, получают в этом случае представление, идентичное, с точностью до констант, представлению предиката ‘будить’ в (10) - (11):

(12) $\lambda e[assemble(e) \wedge agent(basil)(e) \wedge theme(model)(e)]$

(13) $\lambda e[CM(e, \lambda e'.assemble(e') \wedge agent(basil)(e') \wedge \wedge theme(model)(e')]$

В (10) и (12) не содержится никаких эксплицитных ограничений на временные отношения деятельности и изменения состояния. Предположим, что событийный предикат в (10) обозначает события пробуждения, где деятельность непосредственно предшествует изменению состояния. Если предположение верно, почему такая же временная структура невозможна для событийной дескрипции в (12)? Почему не может быть так, что (12) содержит в своем экстенсионале события собирания, в которых вся деятельность агенса предшествует во времени изменению состояния модели? Здравый смысл подсказывает нам, что так собрать модель самолета невозможно, однако (12) не объясняет, почему.

Из этого следует, что распределение подсобытийных компонентов между актуальным и не-актуальными мирами из Таблицы 2 невозможно описать и объяснить, оставаясь в пределах недекомпозиционального подхода к семантике свершений. Различие между НП- и ЧУ-предикатами делается невыразимым: (10) и (12) содержат слишком мало структуры. Необходимым условием для успешного решения этой проблемы оказывается декомпозиция событийных предикатов и задание их подсобытийных компонентов явным образом. Это единственный способ выразить мысль, что в одних случаях подсобытие деятельности размещается в актуальном мире вместе с каузируемым изменением состояния, а в других – отдельно от него. Декомпозициональная теория структуры свершений представлена в следующем разделе.

6.4. ПОДСОБЫТИЯ И ОТНОШЕНИЯ МЕЖДУ НИМИ

Предположим для начала, что интересующие нас предикаты можно анализировать так, как показано в (14) - (15), где целая ситуация подается как сумма каузально связанных деятельности и изменения состояния:

(14) $\lambda e \exists e' \exists e'' [e = e' \oplus e'' \wedge wake_A(e') \wedge agent(basil)(e') \wedge wake_{CS}(e'') \wedge theme(peter)(e'') \wedge cause(e'')(e')]$

где $wake_A$ – предикат, обозначающий деятельности, нацеленные на пробуждение кого-либо, а $wake_{CS}$ – предикат, описывающий изменение состояния пробуждающегося.

(15) $\lambda e \exists e' \exists e'' [e = e' \oplus e'' \wedge assemble_A(e') \wedge agent(basil)(e') \wedge assemble_{CS}(e'') \wedge theme(thread)(e'') \wedge cause(e'')(e')]$

где $assemble_A$ – предикат, обозначающий деятельности по собиранию чего-либо, а $assemble_{CS}$ – предикат, описывающий изменение состояния собираемого объекта.

Сразу же становится ясно, что (14) - (15) дают нам немногое. Помимо дополнительных проблем с каузативным анализом свершений (см. Rothstein 2004: 104; Tatevosov, Ivanov 2009), мы сталкиваемся и с прежней проблемой: ‘будить человека’ и ‘собрать модель’ имеют одинаковую событийную структуру (с точностью до предикатных констант), а различие в некульминирующих прочтениях по-прежнему остаются необъясненными.

Это значит, что раздельное представление деятельности и изменения состояния – условие необходимое, но не достаточное для понимания того, как деривируются НП- и ЧУ-свершения. Чтобы быть уверенными, что их семантические особенности сделаны полностью эксплицитными, необходимо задать *отношение* между двумя подсобытиями.

Большая часть исследователей (см. Таблицу 1) считает, что подсобытия связаны каузально, и именно в этом состоит источник затруднений. Каузальное отношение в (14) - (15) ничего не сообщает нам о характере временной соотнесенности подсобытий (за исключением того тривиального факта, что причина не может иметь место после следствия). Чтобы решить эту проблему, в (Tatevosov, Ivanov 2009) предлагается теория свершений, опирающаяся на теорию С.Ротстин (Rothstein 2004) и расширяющая ее. Теория Ротстин кратко представлена в (16):

(16) (Rothstein 2004): основные определения

a. Событийный шаблон свершений

$$\lambda y \lambda e \exists e_1 \exists e_2 \, [e = {}^{S}(e_1 \cup e_2) \wedge \textit{ACTIVITY}(e_1) \wedge agent(e_1) = x \wedge theme(e_1) = y \wedge \textit{BECOME}(e_2) \wedge arg(e_2) = theme(e_1) \wedge INCR(e_1, e_2, C(e_2))],$$

где ${}^{S}(e_1 \cup e_2)$ – единичная сущность, состоящая из e_1 и e_2

b. Инкрементальное отношение на (под)событиях

Событие e_1 находится в инкрементальном отношении к событию e_2 относительно инкрементальной цепочки $C(e_2)$, $INCR(e_1, e_2, C(e_2))$, тогда и только тогда, когда имеется контекстно-доступная функция μ из $C(e_2)$ в $PART(e_1)$ такая, что $\forall e \in C(e_2)\ \tau(e) = \tau(\mu(e))$.

c. Инкрементальная цепочка $C(e)$.

Инкрементальная цепочка для события $(C(e))$, – это множество частей события e таких, что минимальное событие в $C(e)$ – это

левая граница *e*, для любых e_1, e_2 в $C(e)$ $e_1 \leq e_2$ или $e_2 \leq e_1$, и *e* – также элемент $C(e)$.

Согласно С. Ротстин, свершение есть сумма двух подсобытий, где суммирующая операция ${}^{S}(e_1 \cup e_2)$ создает единую сущность. Релевантными подсобытиями выступают деятельность (e_1 в (16a)) и изменение состояния (e_2 в (16a)). В системе Ротстин подсобытия связаны инкрементально. INCR, инкрементальное отношение, включает в себя контекстно-зависимую функцию, которая устанавливает одно-однозначное соответствие между частями инкрементальной цепочки, (16c), и частями деятельности. Эта функция заменяет обычное каузальное отношение между подсобытиями и гарантирует, что соотнесенные подсобытия происходят одновременно.

Отношение INCR, согласно Ротстин, есть определяющее свойство свершений. В этом ключе она рассматривает глаголы типа ‘шить’, ‘пахать’, ‘читать’, однако не обсуждает предикаты типа ‘рвать нить’, ‘будить человека’ и т.д. Проблема в том, что для таких предикатов отношение между подсобытиями не может быть инкрементальным, поскольку, как мы видели выше, деятельность, вплоть до конечной точки, не вызывает изменения состояния. При НП-прочтении, какая бы деятельность не совершалась, пациенс остается в исходном состоянии.

(Tatevosov, Ivanov 2009) мы выдвинули предположение, что INCR – это есть лишь одно из возможных отношений между подсобытиями деятельности и изменения состояния. НП-предикаты типа ‘будить человека’ имеют ту же событийную структуру, что и ЧУ-предикаты типа ‘собирать модель’, за одним исключением: подсобытия связаны отношением, которое мы называем *отображением в минимальную конечную часть (Mapping to a minimal final part*, MMFP), определение которого дано в (17):

(17) a. Событие e_1 находится в отношении отображения в минимальную конечную часть к событию e_2, $MMFP(e_2)(e_1)$, если и только если имеется контекстно-доступная функция μ из e_2 в $PART(e_1)$ такая, что e_2 отображается в минимальную конечную часть e_1.

b. Событие e' – конечная часть события *e*, $FIN(e)(e')$, если и только если $e' \leq e$ и $\neg\exists e''\,[e'' \leq e \wedge e' \leq e'']$

где “≤” – отношение предшествования на событиях (Krifka 1998: 207).

с. Событие *e′* – минимальная конечная часть e если и только если *FIN(e)(e′)* и $\neg\exists e''\,[FIN(e)(e'') \wedge e'' < e']$.

MMFP – такое отношение, при котором никакая часть подсобытия изменения состояния не отражается в неконечные части подсобытия деятельности. Оно позволяет сделать явным интуитивное представление о том, что неконечные элементы деятельности не вызывают изменений в объекте. Сформулировав разницу между INCR и MMFP, мы можем предложить следующие семантические представления для ЧУ-свершений типа ‘собирать модель’ и НП-свершений типа ‘будить человека’:

(18) $\lambda e\exists e'\exists e''[e = e' \oplus e'' \wedge wake_A(e') \wedge agent(basil)(e) \wedge theme(peter)(e') \wedge wake_{CS}(e'') \wedge arg(e'') = theme(e') \wedge MMFP(e'')(e')]$

(19) $\lambda e\exists e'\exists e''[e = e' \oplus e'' \wedge assemble_A(e') \wedge agent(basil)(e') \wedge theme(model)(e') \wedge assemble_{CS}(e'') \wedge arg(e'') = theme(e') \wedge INCR(e'')(e')(C(e''))]$

Разница между двумя типами свершений сводится, таким образом, к отношению между подсобытиями деятельности и изменения состояния, в полном соответствии с интуитивной трактовкой, предложенной в части 3. ЧУ-свершения опираются на инкрементальное отношение, описанное С. Ротстин, тогда как НП-свершения предполагают отображение в минимальную конечную часть деятельности.

В (18) - (19) представлены кульминирующие варианты НП- и ЧУ-свершений. Некульминирующие, как и прежде, получаются применением оператора CM:

(20) $\lambda e.CM(e, \lambda e_1\exists e'\exists e''[e_1 = e' \oplus e'' \wedge wake_A(e') \wedge agent(basil)(e') \wedge theme(peter)(e') \wedge wake_{CS}(e'') \wedge arg(e'')=theme(e') \wedge MMFP(e'')(e')])$

(21) $\lambda e.CM(e, \lambda e_1\exists e'\exists e''[e_1 = e' \oplus e'' \wedge assemble_A(e') \wedge agent(basil)(e') \wedge theme(model)(e') \wedge assemble_{CS}(e'') \wedge arg(e'')=theme(e') \wedge INCR(e'')(e')(C(e''))])$

(20) обозначает собственные неконечные стадии события, в котором происходит пробуждение пациенса. Они складываются из деятельности, в который Вася задействован как агенс, а Петя как пациенс, и изменения состояния Пети. Собственные неконечные стадии по определению не содержат завершающих частей событий из исходного экстенсионала предиката. Однако благо-

даря MMFP, изменение состояния происходит именно в такой конечной части. Как следствие, предикат в (20) обозначает события, где деятельность агенса не влечет за собой каких-либо изменений состояния. Именно так создается НП-интерпретация.

INCR-свершение ‘собирать модель’ в (21) проходит те же шаги деривации. Его основное отличие от (20) заключается в том, что подсобытие изменения состояния соотнесено с деятельностью инкрементально. Как и (20), (21) обозначает не достигнувшие кульминации ситуации (в данном случае – собирание модели), а их собственные неконечные стадии. Однако вследствие инкрементальности каждая такая стадия влечет за собой некоторое изменение в пациенсе. Так объясняется ЧУ-прочтение ‘собирать’ и других ЧУ-свершений из раздела 2.

Предложенный анализ дает нам существенный эмпирический выигрыш: будучи сочетанием уточненной партитивной теории некульминирования и декомпозициональной теории структуры свершений, он позволяет описать и объяснить в общем виде как сходства, так и различия между НП- и ЧУ-свершениями. Все некульминирующие прочтения возникают с помощью оператора CM, который выводит кульминацию за пределы актуального мира. С другой стороны, допущение о том, что в деривацию свершений вовлечены разные отношения между подсобытиями (INCR vs. MMFP), позволяет зафиксировать различие между НП- и ЧУ-предикатами. Именно эти отношения отвечают за различное распределение подсобытий между нашим миром и мирами, в которых достигается кульминация.

Имея теорию, объясняющую, как выстраиваются некульминирующие прочтения, мы можем вернуться к различию между примерами (a) и (b) - (c) в (6) - (9) и ответить на следующий вопрос: почему некульминирующие прочтения (как ‘неудавшаяся попытка’, так и ‘частичный успех’) возможны не для всех свершений?

6.5. СЕМАНТИЧЕСКИЙ СДВИГ

В предыдущих разделах мы показали, что *если* предикат класса свершений допускает некульминирующее прочтение, оператор CM извлекает собственные неконечные стадии события из его исходного экстенсионала. Однако, (6a) - (9a) показывают, что для свершений типа ‘застрелить человека’ кульминация обязательна.

Интуитивно кажется, что ‘застрелить человека’ похож на MMFP-свершения, рассмотренные выше, такие, как ‘будить человека’: изменение состояния происходит в минимальной конечной точке деятельности. Однако вместо интерпретации с неудавшейся попыткой, эти предикаты вовсе не дают некульминирующих прочтений. Почему? Почему применение оператора CM не приводит к появлению предиката, обозначающего неконечные стадии застреливания человека?

Чтобы ответить на этот вопрос, нам потребуется дополнительное наблюдение: некульминирующие свершение обоих типов (НП и ЧУ) устроены таким образом, что контекстно-идентифицируемые подсобытия, из которых складывается деятельность, не упорядочены ни временным предшествованием, ни каузальной зависимостью.

Предположим, что агенс, пытаясь разбудить пациенса, сначала зовет его шепотом, потом кричит на него, потом хлопает в ладоши у него над ухом, потом трясет его за плечо и, наконец, выливает на него холодную воду. Пациенс просыпается. Эта ситуация описывается кульминирующим предикатом ‘разбудить Петю’ в (3и). Критически важно то, что шепот, крик, хлопание в ладоши и встряхивание за плечо являются частью деятельности агенса и описываются рассматриваемым предикатом. Однако они не являются элементами каузальной цепочки, ведущей к пробуждению. Пробуждение казуально независимо от этих (под)событий, поскольку, если бы они не осуществились, это никак бы не повлияло на пробуждение. Оно все равно бы произошло после выливания воды.

Теперь предположим, что агенс выполняет все эти действия за исключением последнего: как только он видит, что встряхивание Пети за плечо не приводит к пробуждению, он решает прекратить попытки. Это некульминирующее прочтение из (7b); деятельность, протекающая в актуальном мире теперь содержит только подсобытия, от которых изменение состояния каузально независимо. Эти подсобытия каузально независимы и друг от друга: возможно трясти кого-то за плечо, не шепча ему предварительно на ухо, и наоборот. Временная последовательность этих подсобытий также иррелевантна: вне зависимости от того, как они размещаются во времени, их сумма по-прежнему описывается предикатом ‘будить’, то есть входит в его экстенсионал. То же верно и для

‘отрывать пуговицу’, ‘открывать дверь’ и ‘разрывать нитку’ ((6b), (8b) и (9b)).

Из этих наблюдений выстраивается обобщение: при некульминирующей интерпретации (‘неудавшаяся попытка’) MMFP-предикаты типа ‘будить’ описывают деятельность, внутренне не упорядоченную темпоральным предшествованием и/или каузальной зависимостью. Это обобщение распространяется и на компонент деятельности в INCR-предикатах. Как показывает С. Ротстин (Rothstein 2004), лексические значения инкрементальных свершений не предполагают внутренней упорядоченности подсобытия деятельности.

Имеется еще один факт, указывающий на то, что обобщение верно. Модальный оператор CM, согласно гипотезе, – то общее, что есть у некульминирующих свершений и прогрессивов. Первые, однако, обнаруживают ограничения, не присущие вторым. Чтобы увидеть это, обратимся к предикату, лексическое значение которого не фиксирует жестко свойства компонента деятельности в составе сложной событийной дескрипции, например, ‘открывать дверь’ в (4b) и (8b). Во всех исследуемых языках клаузы в прогрессиве, опирающиеся на предикат ‘открывать дверь’, например, (22) тубаларского диалекта алтайского языка, допускает два сценария в (23a-b) (помимо, разумеется, множества прочих):

(22) Vasja eSik-ni aC-yptJit
В. дверь-ACC открывать-IPFV.3SG
Вася открывает дверь.

(23) a. Сценарий 1. Дверь открывается путем введения кода, состоящего из последовательности цифр 1-2-3-4-5-6-7-8. В момент, о котором идет речь, агенс вводит шестую цифру из восьми.

b. Сценарий 2. Замок сломан. Агенс пытается открыть дверь ключом, затем отмычкой, затем ломом, затем пытается выпилить замок и т.д. В момент, о котором идет речь, он совершает одно из этих действий.

В отличие от прогрессива, некульминирующее свершение в (8b) совместимо со сценарием 2, но не со сценарием 1. (8b) возможно, если агенс совершает ряд действий, описанный в (23b), но затем останавливается не добившись результата. Предложение

неприемлемо, если событие прерывается, когда агенс ввел шесть цифр из восьми, как в (23а).

Разница между сценариями 1 и 2, заключается в том, что в первом случае события, ведущие к открытию двери, внутренне упорядочены, а во втором нет. У сценария 1 внутренняя упорядоченность обусловлена последовательностью цифр кода, тогда как у сценарии 2 релевантные элементы деятельности не обязаны совершаться в определенном порядке: они каузально и темпорально не зависимы (можно попытаться выпилить замок, независимо от того, предпринимались ли до того попытки выломать дверь ломом; неважно, совершилась ли попытка вскрыть замок отмычкой до или после попытки открыть дверь ключом).

Из этих наблюдений следует простой вывод: некульминирующие свершения требуют, чтобы деятельность была внутренне неупорядочена. Если это невозможно, они не образуются. Попробуем выразить это обобщение более формально. Мы предполагаем, что отсутствие внутренней упорядоченности сводится к свойству общей гомогенности в (24):

(24) $\forall P[G\text{-}HOM(P) \leftrightarrow T\text{-}HOM(P) \wedge C\text{-}HOM(P)]$

Согласно (24), предикат P общегомогенен, если он темпорально гомогенен, T-HOM(P), и каузально гомогенен, C-HOM(P). Каузальную гомогенность может описать как (25):

(25) $\forall P[C\text{-}HOM(P) \leftrightarrow \forall e[P(e) \rightarrow \forall Q[\forall e'[Q(e') \rightarrow e' \leq e] \rightarrow \forall e''\forall e'''\ [Q(e'') \wedge Q(e''') \rightarrow \neg\ CDEP(e''')(e'')]]]]$

где “≤” мереологическое отношение части и целого, а CDEP – отношение каузальной зависимости.

В соответствии с (24), событийный предикат P каузально гомогенен если для всякого разбиения Q событий из экстенсионала P никакие члены разбиения каузально не зависят друг от друга. (Стандартная точка зрения по вопросу (Lewis 1973) заключается в том, что каузальная зависимость сводима к контрфактической зависимости.)

Замена отношения «каузально зависит от» на «обязательно предшествует» позволяет дать определение темпоральной гомогенности:

(26) $\forall P[T\text{-}HOM(P) \leftrightarrow \forall e[P(e) \rightarrow \forall Q[\forall e'[Q(e') \rightarrow e' \leq e] \rightarrow \forall e''\forall e'''\ [Q(e'') \wedge Q(e''') \rightarrow \neg\ NPREC(e'')(e''')]]]]$

Отношение NPREC можно рассматривать как сочетание метафизической необходимости с темпоральным предшествованием. Мы оставляем прояснение технических деталей до следующего раза.

Итак, некульминирующие свершения содержат в себе общегомогенные деятельности в смысле (24). Семантика в (17) - (18) в текущем виде, однако, этого не учитывает и нуждается в уточнении.

Мы предполагаем, что отсутствие кульминации в мишарском диалекте татарского языка, в чувашском, тубаларском и карачевобалкарском языках есть результат семантического сдвига, преобразующего событийную структуру свершений в структуру деятельностей. Это показано в (27):

(27) $SHIFT_{ACCOMPLSHMENT \rightarrow ACTIVITY}(P) = \lambda e.H(\lambda e'.CM(e', P))(e)$

Вклад оператора $SHIFT_{ACCOMPLSHMENT \rightarrow ACTIVITY}$ в интерпретацию состоит из двух частей. Во-первых, он гарантирует, что событие не кульминирует в актуальном мире, применяя оператор CM к множеству событий из исходного экстенсионала предиката класса свершений P. Во-вторых, он требует, чтобы собственные неконечные стадии P-события, извлеченные оператором CM, были гомогенны в описанном в (24) смысле. Последнее требование реализуется оператором H(omogeneity) логического типа <<v,t,>,<v,t>>, который применяется к событийному предикату и создает новый событийный предикат. Вот возможное определение этого оператора:

(28) $H(P) = \{e \mid P(e) \wedge \exists Q[G\text{-}HOM(Q) \wedge Q \subseteq P \wedge Q(e)]\}$

Согласно (28), результатом применения оператора *H* к *P, H(P)*, является гомогенное подмножество событий из исходного экстенсионала *P*, если в *P* содержится такое подмножество. В противном случае экстенсионал *H(P)* пуст.

С этим уточнением некульминирующий предикат ‘будить Петю’ из (20) превращается в (29). Аналогичное преобразование претерпевает и предикат ‘собирать модель’ в (21).

(29) $SHIFT_{ACCOMPLSHMENT \rightarrow ACTIVITY}(\lambda e \exists e' \exists e''[e = e' \oplus e'' \wedge wake_A(e') \wedge agent(basil)(e) \wedge theme(peter)(e') \wedge wake_{CS}(e'') \wedge arg(e'')=theme(e') \wedge MMFP(e'')(e')]) = \lambda e[H(\lambda e_1.CM(e_1, \lambda e_2 \exists e' \exists e''[e_2 = e' \oplus e'' \wedge wake_A(e') \wedge agent(basil)(e') \wedge theme(peter)(e') \wedge wake_{CS}(e'') \wedge arg(e'')=theme(e') \wedge MMFP(e'')(e')])](e)$

В (29) оператор СМ создает событийный предикат, содержащий собственные неконечные стадии события пробуждения. Как и прежде, поскольку элементы деятельности и изменения состояния связаны отношением MMFP, эти стадии содержат деятельность агенса, но не изменение состояния пациенса. Оператор Н выделяет гомогенное подмножество деятельностей, нацеленных на пробуждение пациенса. Результатом семантического сдвига оказывается не свершение, а деятельность. Читатель может самостоятельно удостовериться, что предикат в (29) удовлетворяет любым разумным ограничениям на семантические характеристики деятельностей.

Теперь мы можем объяснить, почему предикаты типа ‘застрелить человека’ не допускают некульминирующих прочтений. Глаголы класса свершений различаются тем, насколько жестко их лексичяеское значение фиксирует внутреннюю упорядоченность подсобытия деятельности. Деятельности, выступающие компонентом предиката ‘будить человека’ (например, $wake_A$ в (29)), содержат гомогенные подмножества событий. У предикатов типа ‘застрелить человека’ таких подмножеств нет.

Деятельность, которую производит агенс события, описываемого, как ‘застрелить’, состоит из подсобытий (зарядить, передернуть затвор, прицелиться, нажать на спусковой крючок, выстрелить), которые упорядочены как темпорально, так и каузально (например, спускание курка предшествует выстрелу и каузирует его). То же верно и для деятельностей в составе предикатов ‘выключать свет’ или ‘хоронить тело’ в (6а) - (9а). Последовательность подсобытий во всех случаях устроена так, что если действия выполнены в неверном порядке или пропущены, происходящее более не описывается как выключение света или захоронение тела. Иными словами, множества таких деятельностей не способны предоставить оператору Н то, в чем он нуждается: предикат, содержащий гомогенное подмножество. Соответственно, применение оператора Н к таким деятельностям создает пустое множество событий. Такова причина, по которой невозможно некульминирующее прочтение у ‘застрелить человека’ и аналогичных предикатов.

Если это обобщение верно, можно сделать его эксплицитным, приписав соответствующие аксиомы предикатам деятельности, которые являются элементами сложной структуры свершения:

(30) a. $H\text{-}SUBSET(wake_A)$ b. $\neg\, H\text{-}SUBSET(shoot_A)$
c. $\forall P[H\text{-}SUBSET(P) \leftrightarrow \exists Q[Q \subseteq P \wedge G\text{-}HOM(Q)]$

Аксиома в (30a) собщает, что множество событий, обозначаемое предикатом $wake_A$, обладает гомогенным подмножеством. Это гарантирует, что экстенсионал претерпевшего семантический сдвиг свершения в (29) не пуст. Напротив, согласно (30b) событийный предикат $shoot_A$ не имеет гомогенного подмножества, а значит экстенсионал предиката в (31), в прочих отношениях параллельного (29), – пустое множество:

(31) $SHIFT_{ACCOMPLSHMENT \rightarrow ACTIVITY}(\lambda e \exists e' \exists e''[e = e' \oplus e'' \wedge shoot_A(e') \wedge agent(basil)(e) \wedge theme(peter)(e') \wedge shoot_{CS}(e'') \wedge arg(e'')=theme(e') \wedge MMFP(e'')(e')]) =$
$\lambda e[H(\lambda e_1.CM(e_1, \lambda e_2 \exists e' \exists e''[e_2 = e' \oplus e'' \wedge shoot_A(e') \wedge agent(basil)(e') \wedge theme(peter)(e') \wedge shoot_{CS}(e'') \wedge arg(e'') = theme(e') \wedge MMFP(e'')(e')])](e)$

Ограничения на образование некульминирующих свершений, таким образом, получает естественное объяснение. Нам остается подвести основные итоги исследования.

Выводы

Мы выделили три подкласса свершений, различающихся тем, допускают ли они прочтение вида ‘неудавшаяся попытка’, ‘частичный успех’ или ни то ни другое. Опираясь на идеи, независимо обоснованные в литературе, мы предположили, что существенной частью семантической структуры некульминирующих предикатов является модальный оператор, выделяющий из событий их собственные неконечные стадии. Мы показали, что для успешного объяснения свойств некульминирующих свершений необходим их декомпозициональный анализ, когда компоненты деятельности и изменения состояния становятся самостоятельными элементами семантического представления. Разница между НП- и ЧУ-свершениями определяется отношением между подсобытиями в декомпозициональной структуре предиката. Прочтение вида ‘неудавшаяся попытка’ возникает благодаря отношению MMFP (отображение в минимальную конечную часть). ‘Частичный успех’

создается благодаря инкрементальному отношению, описанному в Rothstein 2004. Наконец, мы показали, что некульминирующие свершения обозначают события, у которых компонент деятельности не упорядочен ни темпорально, ни каузально. Это подводит нас к заключительному шагу: предположению, что деривация некульминирующих свершений предполагает семантический сдвиг, превращающий свершения в деятельности. Условие его возможности – темпоральная и каузальная гомогенность возникающего в результате предиката. Это объясняет, почему определенный класс свершений не допускает некульминирующего прочтения: компонент деятельности в их составе негомогенен, то есть внутренне упорядочен.

СПИСОК ЛИТЕРАТУРЫ

Dowty david r. Word meaning and montague grammar. The semantics of verbs and times in generative semantics and in montague's ptq. Series: studies in linguistics and philosophy. – Vol. 7. – Springer netherlands, 1979. – 415 Pp.

Koenig J.-P. and Muansuwan N. How to end without ever finishing: Thai semi-perfectivity // *Journal of Semantics*. – 2001. –Vol. 17. – P. 147–184.

Kratzer A. Building Statives // *Proceedings of the 26th Annual Meeting of the Berkeley Linguistics Society*, 2000. – P. 385–399.

Kratzer A. Building resultatives / C. Maienbaum and A. Wollstein-Leisen (eds.). *Event arguments in Syntax, Semantics, and Discourse*. – Tubingen: Niemeyer, 2005.

Krifka M. The origins of telicity / ed. Susan Rothstein. *Events and Grammar*. – Dordrecht: Kluwer Academic Publishers, 1998. – P. 197– 235.

Landman F. The progressive // *Natural Language Semantics*. – 1992. – Vol. 1.1. – P. 1–32.

Levin B. and Rappaport Hovav M. *Unaccusativity: At the Syntax-Lexical Semantics Interface*, Linguistic Inquiry Monograph 26. – MIT Press, Cambridge, MA, 1995.

Levin B. and Rappaport-Hovav M. *Argument realisation*. – Cambridge: Cambridge University Press, 2005.

Lewis D. Causation // Journal of Philosophy. – 1973. – Vol. 70. – P. 556–567.

Bar-el L., Davis H., and Matthewson L. On Non-Culminating Accomplishments // *Proceedings of the North Eastern Linguistics Society* 35. – Amherst, MA: GLSA., 2005.

Ramchand G.. *Verb Meaning and the Lexicon: A First Phase Syntax*. – Cambridge: Cambridge University Press, 2008.

Rappaport H., Levin M. and B. Building Verb Meanings / M. Butt and W. Geuder (eds.). *The Projection of Arguments: Lexical and Compositional Factors.* – CSLI Publications, Stanford, CA, 1998. – P. 97–134.

Rothstein S. *Structuring events: a study in the semantics of lexical aspect*. – Malden (Mass.): Blackwell publishing, 2004.

Tatevosov S. and Ivanov M. Event structure of non-culminating accomplishments / Hogeweg, Lotte, Helen de Hoop, and Andrej Malchukov (eds.). *Cross-linguistics Semantics of Tense, Aspect, and Modality.* – Amsterdam: John Benjamin, 2009. – P. 83–130.

ПРАГМАТИЧЕСКИ ОРИЕНТИРОВАННЫЕ ЛИНГВИСТИЧЕСКИЕ МОДЕЛИ КАК ОСНОВА СИСТЕМ И ТЕХНОЛОГИЙ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА. АНАЛИТИЧЕСКИЙ ОБЗОР

Д. Ш. Сулейманов

Институт прикладной семиотики Академи наук РТ,
Казанский федеральный университет

В работе представлен анализ существующих лингвистических моделей, систем и информационных технологий обработки ЕЯ-текстов; исследованы современные подходы и методы построения лингвистических моделей, систем и информационных технологий обработки естественно-языковых текстов, а также определены принципы и критерии их построения.

Ключевые слова: Семиотическая модель, прагматически-ориентированная модель, диалоговая система, вопросно-ответная система, обработка естественного языка.

ОНТОЛОГО-ЛИНГВИСТИЧЕСКИЕ СИСТЕМЫ: БАЗОВЫЕ МОДЕЛИ

О. А. Невзорова

Институт прикладной семиотики Академии наук РТ

Онтолого-лингвистические системы ориентированы на решение сложных задач обработки естественного языка, требующих семантических знаний. В основе проектирования онтолингвистических систем лежат процессы скоординированного взаимодействия различных уровней, прежде всего онтологического (обеспечивающего системные модели знаний о мире) и различных языковых уровней. В работе рассматриваются концептуальная архитектура онтолого-лингвистических систем, а также решения по организации системы онтологических моделей в системе «OntoIntegrator».

Ключевые слова: Онтолого-лингвистическая система, онтологическая модель, лингвистическая онтология, онтология, обработка естественного языка.

СОЗДАНИЕ ПЕРСОНАЛЬНОЙ ОНТОЛОГИИ

П. И. Соснин

Ульяновский государственный технический университет

В работе приводятся результаты исследований вопросов онтологизации персонального профессионального опыта. Специфику решений по созданию и использование персональной онтологии определяет ориентация на представление опыта в виде системы моделей прецедентов, освоенных и созданных субъектом деятельности в реализации совокупности проектов. Модели прецедентов построены по образцу интеллектуально обработанных «деятельностных рефлексов» с использованием вопросно-ответных рассуждений.

Ключевые слова: Онтология, персональная онтология, вопросно-ответная память, персональный опыт, структура данных.

ЛЕКСИКОГРАФИЧЕСКИЕ РЕСУРСЫ ПЕРЕВОДЧИКА: СОСТАВ, СТРУКТУРА И ВЕДЕНИЕ

Л. Н. Беляева

Российский государственный педагогический университет им. А. И. Герцена

Современное исследование в области создания лексикографических ресурсов предполагает проведение предварительной терминологической работы для отбора и описания терминологии на разных языках, осуществление гармонизации этих описаний и согласование терминологических систем разных языков. Лексикографические ресурсы переводчика определяют оперативность, точность и корректность результатов его работы. Рассматривается исторический аспект разработки таких ресурсов, современные сетевые базы данных и методы их поддержки и ведения. Отдельно анализируются возможности использования лексикографических ресурсов в образовательной среде вуза.

Ключевые слова: Лексикографический ресурс, перевод, терминология, терминологический банк данных, извлечение терминов.

ОБЪЕКТНО-ОРИЕНТИРОВАННЫЙ АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТОВ

Н. В. Лукашевич

Московский государственный университет им. М. В. Ломоносова

Работа посвящена рассмотрению подходов к анализу тональности текстов по отношению к заданному объекту, а также его характеристикам (аспектам). Для решения задачи анализа тональности по отношению к аспектам сущности необходимо решать также задачи извлечения аспектов для сущности, категоризацию или кластеризацию аспектов по аспектным категориям, определение тональности текста по отношению к заданному аспекту или аспектной категории. Также в работе описываются открытые тестирования объектно-ориентированных систем анализа тональности.

Ключевые слова: Анализ тональности, аспектный термин, частотность, оценочное слово.

ЗАМЕТКИ О ТЮРКСКИХ СВЕРШЕНИЯХ

С. Г. Татевосов

Московский государственный университет имени М. В. Ломоносова

В работе рассмотрены некульминирующие прочтения предикатов класса свершений. Показано, что для объяснения свойств некульминирующих свершений необходим их декомпозиционный анализ. Различные глаголы по-разному ограничивают отношения между подсобытиями; свершения различаются с точки зрения внутренней упорядоченности подсобытия.

Ключевые слова: Тюркский язык, предикат, свершение, некульминирующее свершение, структура события.

СВЕДЕНИЯ ОБ АВТОРАХ

Сулейманов Джавдет Шевкетович – академик АН РТ, доктор технических наук, профессор, директор Института прикладной семиотики АН РТ, заведующий кафедрой информационных систем Казанского федерального университета (Казань).

Область научных интересов: искусственный интеллект, компьютерная лингвистика, компьютерные технологии в образовании.

Невзорова Ольга Авенировна – кандидат технических наук, доцент, заместитель директора Института прикладной семиотики АН РТ (Казань).

Область научных интересов: онтологическое моделирование, искусственный интеллект, математическая лингвистика, компьютерная лексикография.

Соснин Петр Иванович – доктор технических наук, профессор, заведующий кафедрой вычислительной техники Ульяновского государственного технического университета (Ульяновск).

Область научных интересов: искусственный интеллект, вопросно-ответные процессы, технологии и системы, онтологическое моделирование.

Беляева Лариса Николаевна – доктор филологических наук, профессор, заведующий кафедрой образовательных технологий в филологии филологического факультета Российского государственного педагогического университета имени А. И. Герцена (Санкт-Петербург).

Область научных интересов: математическая лингвистика, компьютерная лексикография, переводоведение, высокотехнологическая образовательная среда.

Лукашевич Наталья Валентиновна – доктор технических наук, ведущий научный сотрудник Лаборатории анализа информационных ресурсов Научно-исследовательского вычислительного центра МГУ им. Ломоносова (Москва).

Область научных интересов: онтологическое моделирование, математическая лингвистика, компьютерная лексикография.

Татевосов Сергей Георгиевич – доктор филологических наук, профессор кафедры теоретической и прикладной лингвистики филологического факультета МГУ им. М. Ломоносова (Москва).

Область научных интересов: семантика, морфология, типология, языки Кавказа, тюркские языки, уральские языки.

ОГЛАВЛЕНИЕ

Предисловие 3

ГЛАВА 1. ПРАГМАТИЧЕСКИ-ОРИЕНТИРОВАННЫЕ ЛИНГВИСТИЧЕСКИЕ МОДЕЛИ КАК ОСНОВА СИСТЕМ И ТЕХНОЛОГИЙ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА. АНАЛИТИЧЕСКИЙ ОБЗОР 8
1.1. Введение 8
1.2. Прагматически-ориентированный подход к разработке лингвистических моделей 11
1.3. Анализ близких идей и подходов 21
1.4. Анализ средств обработки ЕЯ-текстов в диалоговых системах (в аспекте прагматически-ориентированного подхода) 24
1.5. Формализмы в основе систем семантической обработки ЕЯ-текстов 31
1.6. Объектно-предикатная система как составляющая концептуально-функциональной модели 37
1.7. Анализ средств формального описания значений 39
1.8. Анализ систем обработки ЕЯ-текстов на основе концептуально-формальной модели 46

ГЛАВА 2. ОНТОЛОГО-ЛИНГВИСТИЧЕСКИЕ СИСТЕМЫ: БАЗОВЫЕ МОДЕЛИ 60
2.1. Базовые предпосылки проектирования онтолого-лингвистических систем 60
2.2. Концептуальная архитектура онтолого-лингвистических систем 62
2.3. Система онтологических моделей как ядро интеллектуальных систем 67
2.4. Онтология планирования задач 69
2.5. Прикладные лингвистические онтологии в онтолого-лингвистической системе 71
2.6. Модель лингвистической оболочки онтологии 75

ГЛАВА 3. СОЗДАНИЕ ПЕРСОНАЛЬНОЙ ОНТОЛОГИИ 84
3.1. Введение 84
3.2. Предварительные основы 85

3.3. Вопросно-ответная память 86
3.4. Онтологизация персонального опыта 90
3.5. Средства создания персональной онтологии 93
3.6. Реализация структур данных 98
3.7. Использование персональной онтологии 99
3.8. Информационные источники 100
3.9. Проектирование конфигурируемых шаблонов 101

ГЛАВА 4. ЛЕКСИКОГРАФИЧЕСКИЕ РЕСУРСЫ ПЕРЕВОДЧИКА: СОСТАВ, СТРУКТУРА И ВЕДЕНИЕ 106
4.1. Введение 106
4.2. Особенности лексикографических ресурсов переводчика . . . 107
4.3. Терминологические банки данных как сетевой лексикографический ресурс 109
4.4. Автоматизация процедуры извлечения терминов из текстов на разных языках 120
4.5. Лексикографический ресурс в составе образовательной среды вуза 126

ГЛАВА 5. ОБЪЕКТНО-ОРИЕНТИРОВАННЫЙ АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТОВ 132
5.1. Введение 132
5.2. Классификация аспектных терминов 133
5.3. Автоматизация выявления признаков/свойств товаров/услуг 136
5.4. Определение тональности по отношению к аспектным терминам 145
5.5. Тестирование объектно-ориентированных система анализа тональности текстов 149

ГЛАВА 6. ЗАМЕТКИ О ТЮРКСКИХ СВЕРШЕНИЯХ 161
6.1. Предикаты класса свершений: теоретические альтернативы 161
6.2. Некульминирующие свершения в тюркских языках 163
6.3. Кульминация и ее недостижение 166
6.4. Подсобытия и отношения между ними 169
6.5. Семантический сдвиг 173

Аннотации к главам 182
Сведения об авторах 185

ФОРМАЛЬНЫЕ МОДЕЛИ
И СИСТЕМЫ В ВЫЧИСЛИТЕЛЬНОЙ
ЛИНГВИСТИКЕ

Подписано в печать 16.11.2016. Формат 60×84 1/16.
Печать офсетная. Гарнитура «TimesNewRoman».
Усл.-печ. л. 10,9. Тираж 000 экз. Заказ

Издательство Академии наук
Республики Татарстан
420111, г. Казань, ул. Баумана, 20
e-mail: izdat.anrt@yandex.ru