Глава 5
АНАЛИЗ ВОЗМОЖНОСТЕЙ ПРИМЕНЕНИЯ АППАРАТА СК-ЯЗЫКОВ К РЕШЕНИЮ РЯДА АКТУАЛЬНЫХ ПРОБЛЕМ ИНФОРМАТИКИ

Данная глава посвящена исследованию возможностей применения аппарата стандартных К-языков (СК-языков) к разработке языков представления содержания посланий компьютерных интеллектуальных агентов, языков формирования контрактов и протоколов переговоров в области электронной коммерции, созданию семантического сетевого языка нового поколения, построению онтологий предметных областей, разработке новых языков представления знаний для решения информационно-сложных задач, проектированию интеллектуальных информационно-поисковых и вопросо-ответных Интернет-систем нового поколения. Содержание данной главы базируется на публикациях (Фомичев 1991, 1992б, 2001а, 2002б, 2005а – 2005в, 2005д; Fomichov 1998b - 2001, 2002b, 2004, 2005).

5.1. Определение класса стандартных К-языков как формальная метаграмматика для описания содержания посланий компьютерных интеллектуальных агентов

5.1.1. Проблема разработки языков общения компьютерных интеллектуальных агентов

Прогресс исследований в области искусственного интеллекта (ИИ) и компьютерных сетей привел к появлению теории многоагентных систем (МАС). Многоагентные системы являются одним из наиболее быстро развивающихся в 1990-е годы и начале 2000-х годов направлений информатики.

Главная причина постоянно растущего интереса к этому направлению заключается в следующем. В настоящее время можно прогнозировать бурное развитие в ближайшие годы *электронной коммерции* (electronic commerce, E-commerce), базирующейся на широчайших возможностях сети Интернет (Thome и Schihzer 1998); МАС рассматриваются как ключевая технология для конструирования систем электронной коммерции (Guilfoyle, Jeffcoate, Stark 1997; Wooldridge 1998).

В частности, одной из перспективных областей использования МАС является индустрия туристического сервиса. Задача проектирования компьютерных интеллектуальных агентов (КИА) для этой индустрии является весьма естественной, поскольку различные КИА многих поставщиков услуг (заказ авиабилетов и железнодорожных билетов, резервирование комнат в гостиницах, взятие напрокат автомобилей, организация культурной программы и т.д.) должны динамически обнаруживать друг друга и эффективно взаимодействовать для решения стоящих перед ними задач.

Многочисленные КИА (сконструированные разными научно-исследовательскими центрами, использующие различную аппаратуру и программное обеспечение) смогут эффективно взаимодействовать в процессе решения своих задач только в том случае, когда эти КИА будут располагать общим языком для обмена информацией и руководствоваться едиными правилами общения. Поэтому в 1990-е годы было разработано несколько *языков общения интеллектуальных агентов*; два из них являются наиболее широко применимыми.

Во-первых, это язык KQML, разработанный в США в рамках проекта разделения знаний, осуществлявшегося национальным агентством ДАРПА (DARPA). Исследования по разработке этого языка отражены, в частности, в публикациях (Finin и др. 1993; Labrou 1996; Finin, Labrou, Mayfield 1997; Labrou, Finin 1997, 1998). Значительную роль в этих исследованиях сыграли специалисты Стэндфордского университета.

Второй язык разработан в рамках международного Фонда интеллектуальных физических агентов, или ФИФА (the Foundation for Intelligent Physical Agents, или

FIPA), штаб-квартира которого находится в Женеве. Одним из важных результатов исследований, организованных этим фондом, стала разработка в 1997 - 1999 годах стандарта для представления посланий (messages) КИА, получившего название Языка общения агентов (FIPA Agent Communication Language = FIPA ACL) (FIPA 1998a). Теоретической основой для создания этого языка послужили принципы разработки языка KQML и языка KIF, или Knowledge Interchange Format (Genesereth, Fikes и др., 1992; Genesereth, 1999). Та часть языка FIPA ACL, которая предназначена для представления содержания посланий КИА, называется *семантическим языком* (FIPA Semantic Language = FIPA SL).

Проблема создания адекватных логико-информационных основ электронной коммерции предъявляет высокие требования к языку представления содержания посланий КИА. Этот язык должен позволять отображать содержание коммерческих переговоров. Однако язык FIPA SL обладает многими ограничениями в этом отношении. В связи с этим возникает проблема разработки математического описания такого класса языков, который был бы удобен для отображения содержания произвольных посланий КИА.

В нашей стране в последнее десятилетие проблематике многоагентных систем уделялось значительное внимание многими учеными. В частности, различные аспекты теории МАС и вопросы применения этой теории исследовались в работах В.И. Городецкого (Городецкий 1998), В.В. Емельянова (Emelyanov 2001), Э.С. Клышинского (Клышинский 1999), Г.С. Плесневича и В.Б. Тарасова (Тарасов 1998;.Plesniewicz, Tarassov 2001), Д.А. Поспелова (Поспелов 1998), Г.В. Рыбиной и В.Ю. Берзина (Рыбина, Берзин 2002).

Однако следует отметить, что проблематика разработки формальных языков с широкими выразительными возможностями для представления содержания посланий КИА не получила в трудах ученых нашей страны такого же внимания, как и в серии зарубежных проектов, выполненных в рамках международного Фонда интеллектуальных физических агентов.

5.1.2. Возможности стандартных К-языков для представления содержания посланий компьютерных интеллектуальных агентов

СК-языки обладают целым рядом свойств, делающих их удобными для представления содержания произвольных посланий КИА. Рассмотрим некоторые из этих свойств.

Свойство 1. К-языки позволяют строить формальные составные обозначения понятий.

Пример 1. Пусть П1 = “тургруппа, состоящая из 12 ученых”, тогда возможным К-представлением (КП) выражения П1 является цепочка

*тур-группа * (Колич-элементов, 12) (Качеств-состав, ученый)* .

Пример 2. Если П2 - выражение “керамика, выпущенная в Индии или Шри-Ланке”, то возможно первое КП П2: *керамика1 * (Производство, (Индия V Шри-Ланка))* и второе КП *керамика1 * (Производство, (нек страна * (Назв, ‘Индия’) : x1 ∨ нек страна * (Назв, ‘Шри-Ланка’) : x2))* .

Пример 3. Пусть П3 = “контейнер, содержащий 8 коробок с чайными сервизами из Китая и 4 коробки со столовыми сервизами из Индии или Шри-Ланки”. Тогда найдется такой концептуальный базис *B*, что *Ls (B)* включает следующее выражение, являющиеся возможным семантическим представлением П3:

*контейнер1 * (Содержание1, (нек множество * (Колич-элементов, 8)*
*(Кач-состав, коробка1 * (Содержание2, сервиз 1 * (Вид, чайный)*
*(Страна, нек страна1 * (Название, ‘Китай’) : x1))) : S1 ∧*
*нек множество * (Колич-элементов, 4)*
*(Кач-состав, коробка1 * (Содержание2, сервиз 1 * (Вид, столовый)*
*(Страна, нек страна 1 * (Название, “Индия” V “Шри-Ланка”) : x2))) : S2))* .

Свойство 2. СК-языки предоставляют широкие возможности построения определений понятий в виде (а) $(c \equiv b * (r_1,d_1)\ldots(r_n,d_n))$

или (б) $((c \equiv b * (r_1,d_1)\ldots(r_n,d_n)) \wedge A)$,

где c – вводимое понятие, b – базовое понятие (считается известным), $n \geq 1$, $r_1,\ldots, r_n$

бинарные реляционные символы; d_1 ,…, d_n – К-цепочки, A - К-цепочка, интерпретируемая как высказывание о свойствах объектов, характеризуемых понятием с..

Пример 1. Пусть П1 – определение “Freight forward – это груз, оплачиваемый в порту назначения (англ. язык)”. Тогда СП определения П1 может являться следующим выражением некоторого СК-языка:

*(freight-forward ≡ груз1 * (Описание1, < x1, Оплата (x1, нек порт 1 * (Пункт-назначения, x1)) >) (Язык, английский))* .

Пример 2. Пусть П2 = “Малое предприятие – это предприятие с количеством сотрудников, не превышающим 50 человек”. Тогда возможным КП определения П2 является следующее выражение:

Определение1 (Малое предприятие, x1, (Явл1 (x1, предприятие) ∧ ¬ Больше (Колич-элементов (Штат 1 (1x)), 50))) .

Здесь используется другая форма представления определения:

Определение1 (c, v, Des (v)) ,

где *c* – обозначение понятия (поясняемого), *v* – переменная (обозначает произвольную сущность, характеризуемую понятием *c*), *Des (v)* – К-цепочка (или *l*-формула), являющаяся СП высказывания о сущности с меткой *v*.

Свойство 3. С помощью СК-языков можно строить составные обозначения различных сущностей, в том числе обозначения множеств, для этого сначала строится составное обозначение понятия (см. свойство 1), причем применяется правило Р[8], а затем добавляется квантор референтности с помощью правила Р[1].

Пример 1. К-цепочка *коробка1 * (Содержание2, сервиз1 * (Вид, чайный))* , построенная в результате применения на последнем шаге правила Р[8] , интерпретируется как составное обозначение понятия “коробка с чайными сервизами”. В результате применения правила Р[1] можно получить выражения

нек коробка 1 (Содержание2, сервиз1 * (Вид, чайный))* , (*)

*все коробки1 * (Содержание2, сервиз1 * (Вид, чайный))* . (**)

Выражение (*) будем интерпретировать как обозначение коробки, где находится один или несколько чайных сервизов. Выражение (**) будем рассматривать как обозначение множества, состоящего из всех коробок, содержащих чайные сервизы.

Пример 2. Мы можем следующим образом обозначить конкретную запланированную серию из 5-ти поставок, каждая из которых включает 60 чайных сервизов № 53 и 36 столовых сервизов № 65:

*Нек множество * (Колич-элементов, 5) (Кач-состав, поставка1 * (Состав2, (нек множество * (Колич-элементов, 60) (Кач-состав, сервиз1 * (Вид, чайный) (Номер, 53)) ∧ нек множество * ((Колич-элементов, 36)*

*(Кач-состав, сервиз1 * (Вид, столовый) (Номер, 65))))) : S1* .

Свойство 4. Следствием свойства 3 является возможность моделировать способ использования йота-оператора в языке FIPA ACL, он включает, в частности, выражение *(iota ? x (UK Prime Minister ? x)* , интерпретируемое как семантическое представление выражения “премьер-министром Великобритании”. Тогда можно построить эквивалентное К-представление этого выражения

*нек чел * (Премьер-министр, UK) : x1* .

Для построения этого КП нужно выполнить следующие шаги:

(1) с помощью правил P[0], P[1] , P[4] построить цепочку

Премьер-министр (нек чел, UK);

(2) по правилу P[8] получить цепочку *чел * (Премьер-министр, UK);*

(3) по правилу P[1] слева приписывается квантор референтности *нек*;

(4) по правилу P[5] справа приписывается цепочка *: x1.*

Свойство 5. СК-языки удобны для построения семантических представлений простых и составных целей.

Пример 1. Пусть G1 – цель “Зарезервировать 12 одноместных номеров в трёх-звёздочных отелях Любляны.”. Тогда К-представлением этой цели может являться выражение

*Резервир1 * (Объект1, нек .множество * (Колич-элементов, 12) (Кач-состав, номер1 * (Вид1, одноместн) (Место1, произвольн отель * (Вид2, трёхзвёздочн) (Локализация, нек город * (Назв, ‘Любляна’) : x1))) : S1) : goal1.*

Пример 2. Пусть G2 = “Доставить фирме “Spencer & Co” в течение 12-19 февраля 2004 года 5 партий, каждая из которых состоит из 60 чайных сервизов № 53 и 56 столовых сервизов № 65”. Тогда цель (или распоряжение) G2 может иметь следующее К-представление:

*Доставка1 * (Адресат1, нек фирма1 * (Назв, ‘Spencer & Co’) : x1)*

(Время, <12.02.2004, 19.02.2004>) (Объект1, setdescr1),

где *setdescr1* – построенное выше КП выражения “серия из 5-ти поставок, каждая из включает 60 чайных сервизов № 53 и 36 столовых сервизов № 65.”

Свойство 6. СК-языки дают возможность представлять коммуникативные акты языка FIPA ACL. Предположим, что компьютерный интеллектуальный агент (КИА) “Клиент-агент” просит КИА “Антологический агент” (электронный поставщик знаний) сообщить, какие бывают виды цитрусовых. Тогда этот коммуникативный акт можно представить в виде следующего выражения некоторого СК-языка:

*нек коммуник-акт * (Вид, запрос)(Отправитель, Клиент-агент)*

*(Получатель, Онтол-агент) (Содержание1, (Вопрос (x1 ≡ нек множество **

*(Кач-состав, произв понятие * (Конкретизация, цитрус)))))*

(Язык-запроса, СК-язык)(Онтология, fipa-ontol-service- fruits-ontology)

(Метка-ответа, цитрус-запрос) (Язык-ответа, СК-язык).

Свойство 7. СК-языки удобны для построения СП вопросов.

Пример. Пусть В1 = “Сколько стоит двухтонный контейнер?”. Тогда возможное К-представление В1:

*Вопрос (x1, (x1 ≡ Цена (произв контейнер1 * (Вес, 2/ тонна)))).*

Свойство 8. СК-языки позволяют использовать ту же форму для построения общих вопросов, т.е. вопросов с ответом ДА / НЕТ.

Пример. Пусть В2 = “Работает ли Сергей Сомов в фирме IBM?”

Тогда К-представление В2 может являться следующим выражением:

*Вопрос (x1, (x1 ≡ Ист-знач. ((∃e1 (ситуация) Явл1 (e1, работа1 **

*(Агент1, нек человек * (Имя, ‘Сергей’) (Фамилия, ‘Сомов’) : x2) (Место3,*

*нек фирма * (Назв, ‘IBM’) : x3)) ∧ Время (e1, #сейчас#))))) .*

5.2. Анализ возможностей использования СК-языков для формирования контрактов и протоколов переговоров в области электронной коммерции

В течение нескольких последних лет в области электронной коммерции возникли два взаимосвязанных научных направления, получивших названия *электронные переговоры (e-negotiations)* и *электронное заключение контрактов (electronic contracting)*. Рождение этих направлений было формально обозначено проведением в начале 2000-х годов нескольких международных конференций и симпозиумов, в том числе конференции по электронным контрактам и вычислениям, базирующимся на контрактах (Цюрих, Швейцария, 2001); 6-й международной конференции по информационным системам для бизнеса (Колорадо Спрингс, США, 2003); симпозиума по теории и применениям электронных переговоров (Познань, Польша, апрель 2004); 1-го международного симпозиума по электронному заключению контрактов, организованного Международным институтом инженеров по электричеству и электронике (IEEE) в июле 2004 г. в Сан Диего, Калифорния, США.

К центральным задачам, стоящим перед исследователями в этих научных направлениях, относится создание формальных языков для представления содержания коммерческих переговоров, проводимых компьютерными интеллектуальными агентами (КИА), и для построения контрактов, заключаемых КИА в ходе таких переговоров. Эти задачи можно рассматривать как важные частные случаи проблемы разработки формальных языков общего назначения для бизнес-коммуникаций (Kimbrough и Moore 1997; Hasselberg и Weigand 2001).

В работе (Hasselberg и Weigand 2001) подчеркивается, что если послания в области электронной коммерции дожны обрабатываться автоматически, то значения (meanings) посланий должны быть формализованы. Эта идея совпадает с высказанным в статье (Kimbrough и Moore 1997) мнением о необходимости

создания логико-семантических основ конструирования формальных языков для бизнес-коммуникаций (ФЯБК).

Кимбру и Мур в указанной выше работе предлагают использовать как можно шире аппарат логики первого порядка для построения выражений любого ФЯБК. Однако выразительные возможности класса языков логики первого порядка очень ограничены с точки зрения описания семантической структуры произвольных бизнес-документов.

Анализ показывает, что протоколы коммерческих переговоров и контрактов могут формироваться с помощью выразительных механизмов естественного языка (ЕЯ), используемых для построения произвольных ЕЯ-текстов, относящихся к медицине, технике, юриспруденции и т.д. В частности, тексты из таких документов могут включать: (а) неопределенные формы глаголов (или инфинитивы) с зависимыми словами, выражающие цели, предложения (“продать 50 ящиков с яблоками”), обещания, обязательства и назначения предметов; (б) конструкции, образованные из инфинитивов с зависимыми словами с помощью логических связок “и”, “или”, “не” и являющиеся составными обозначениями целей, предложений, обещаний, обязательств и назначений предметов; (в) составные обозначения множеств (“партия, состоящая из 50 ящиков с яблоками”); (г) фрагментов, в которых логические связки “и”, “или” соединяют не обозначения высказываний, а обозначения предметов; (д) фрагментов, содержащих ссылки на смысл фраз или более крупных фрагментов дискурса (“это предложение”, “его распоряжение”, “это обещание” и т.д.); (е) обозначения функций, аргументами и/или значениями которых могут быть множества объектов (“персонал фирмы А”, “поставщики фирмы А”, “количество поставщиков фирмы А”); (ж) вопросы с ответом “Да” или “Нет”; (з) вопросы с вопросительными словами.

Между тем, логика первого порядка не позволяет строить формальные аналоги (на семантическом уровне) таких текстов из бизнес-документов, для которых выполняются перечисленные выше свойства (а) – (ж).

Поэтому проблема разработки формальных языков, позволяющих отображать содержание протоколов коммерческих переговоров, проводимых КИА, и

формировать контракты, заключаемые в ходе таких переговоров, очень сложна. В этой связи представляется разумным использовать для решения этой проблемы наиболее широко применимые теории представления структурированных значений ЕЯ-текстов, предоставляемые математической лингвистикой и математической информатикой.

СК-языки обладают рядом выразительных возможностей, необходимых для формального представления содержания контрактов. Для иллюстрации важной части таких возможностей рассмотрим сценарий взаимодействия деловых партнеров в ходе обработки страховой компанией поступившего заявления о повреждении автомобиля. Этот сценарий опубликован в работе (Xu, Jeusfeld 2003), посвященной электронным контрактам. Деловыми партнерами являются страховая компания “AGFIL”, фирмы с названиями “Europ Assist”, “Lee Consulting Services”(сокращенно “Lee C.S.”), а также сервисные центры и техничские эксперты. Фирма “Europ Assist” предоставляет владельцам страховых полисов возможность круглосуточного (24 часа) экстренного обращения по телефону. Фирма “Lee C.S.” ежедневно координирует и управляет операциями срочного обслуживания по соглашению с компанией “AGFIL”.

В целом процесс страхового обслуживания осуществляется следующим образом. Владелец страхового полиса звонит в “Europ Assist”, чтобы заявить о новом страховом случае. Фирма “Europ Assist” регистрирует поступившую заявку, предлагает подходящий сервисный центр и направляет извещение компании “AGFIL”, которая проверяет действительность полиса и то, покрывает ли полис заявку. После того, как компания “AGFIL” получает заявку, эта компания направляет детали заявки фирме “Lee C.S.”.

Компания “AGFIL” посылает владельцу полиса письмо, содержащее полную форму заявления. Фирма “Lee C.S.” соглашается с затратами на ремонт, если эксперт не требуется в связи с тем, что размер ущерба мал; в противном же случае назначается эксперт. Эксперт осматривает поврежденный автомобиль и договаривается с сервисным центром о затратах на ремонт. После получения от фирмы “Lee C.S.” договора о ремонте автомобиля сервисный центр начинает

ремонт. После завершения ремонта сервисный центр посылает счет фирме “Lee C.S.”, которая сравнивает счет с первоначальной оценкой. Фирма “Lee C.S.” возвращает все счета компании “AGFIL”. Эта компания осуществляет выплаты. Если заявка будет признана недействительной, то об этом будут проинформированы все стороны, участвующие в процессе, и процесс будут остановлен.

Данный сценарий позволяет проиллюстрировать ряд свойств СК-языков, делающих их удобным инструментом формального описания контрактов.

Свойство 1. Возможность построения составных обозначений целей.

Пример. Пусть Т1 = “Владелец полиса звонит в фирму “Europ Assist”, чтобы сообщить о повреждении автомобиля”. Тогда Т1 может иметь К-представление (КП)

*Ситуация (e1, телеф-разговор * (Агент1, нек чел * (Владеет1, нек полис1))(Объект2, нек фирма * (Назв, “Europ Assist”)(Цель, сообщение1 * (Тема1, нек повреждение1 * (Объект1, нек автомобиль))))*.

Свойство 2. Наличие средств компактного представления временных и причинно-следственных отношений между ситуациями.

Свойство 3. Возможность построения компактных семантических образов таких фрагментов предложений, которые получены в результате соединения логическими связками “И”, “ИЛИ” обозначений предметов, событий, понятий или целей.

Пример. Пусть Т2 = “После получения накладной по ремонту от фирмы “Lee C.S.” и заявления от владельца страхового полиса, компания “AGFIL” оплатит сервисному центру стоимость ремонта”. Тогда КП текста Т2 может являться выражением

*(Ситуация (e1, (получение1 * (Агент2, нек фирма * (Назв, “AFGIL”) : x1)(Объект1, нек накладная * (Тема, нек ремонт : e2) : x2)(Отправитель, нек фирма * (Назв, “Lee C.S.”) : x3) ∧ получение1 * (Агент2, x1)(Объект1, нек заявление1 : x4) (Отправитель, нек чел * (Владеет1, нек полис1 : x5) : x6))) ∧ Ситуация (e2, оплата1* (Агент2, x1)(Адресат1, нек сервис-центр : x7)(Сумма, Стоимость (e2))) ∧ Раньше (e1, e2))* .

Свойство 4. Существование средств формального представления содержания дискурсов со ссылками на смысл фраз и более крупных фрагментов текста.

Пример. Пусть Т3 = “Фирма “Europ Assist” предоставляет телефонное обслуживание владельцу страхового полиса; в частности, указывает сервисный центр для ремонта и извещает компанию “AGFIL” о заявлении владельца страхового полиса”. Тогда Т3 может иметь КП

*(Ситуация (e1, обслуживание1 * (Агент2, нек фирма * (Назв, “Europ Assist”) : x1)(Инструмент, нек телефон : x2)(Объект1, произвольн чел * (Владеет1, нек полис1 : x3) : x4)) : P1 ∧ Конкретизация (P1, (Ситуация (e2, указание1* (Агент2, x1)(Адресат, x4))(Объект3, нек сервис-центр * (Назначение1, ремонт) : x5)) ∧ Ситуация (e3, извещение1* (Агент2, x1)(Адресат1, нек фирма * (Назв, “AFGIL”) : x6)(Содержание1, нек заявление1* (Автор, x4) : x7))))) .*

Свойство 5. Возможность формального представления содержания контрактных обязательств, зависящих от условий.

Пример. Пусть Т4 = “Фирма “Lee C.S.” назначает эксперта для осмотра автомобиля в течение 41 часа с момента получения заявления о повреждении автомобиля, если стоимость ремонта не превышает 500 USD”. Тогда КП текста Т4 может являться выражением

*Если-то (¬ Больше1 (Стоимость (нек ремонт * (Объект1, нек автомобиль : x1) : e1), < 500, USD >) , (Ситуация (e2, назначение1* (Агент2, нек фирма * (Назв, “Lee C.S.”) : x2)(Персона1, нек эксперт : x3)(Цель1, нек осмотр * (Объект1, x1) : e3)(Момент, t1)) ∧ ¬ Больше1 (Разность (t1, t0), <41, час >) ∧ Ситуация (e4, получение1 * (Агент2, x2)(Объект1, нек заявление1* (Тема, нек повреждение1 * (Объект1, x1) : e5))(Время, t0)))) .*

Выходя за рамки обсуждавшегося в данном параграфе сценария взаимодействия деловых партнеров в процессе обслуживания заявления о страховом случае, сформулируем два дополнительных свойства СК-языков.

Свойство 6. Наличие средств построения составных обозначений множеств как компонентов семантических представлений ЕЯ-текстов, являющихся протоколами переговоров или контрактами (см. в предыдущем параграфе

обозначение запланированной серии из 5-ти поставок, каждая из которых включает 60 чайных сервизов № 53 и 36 столовых сервизов № 65).

Свойство 7. Возможность построения объектно-ориентированных СП протоколов переговоров и контрактов, т.е. формальных выражений вида

нек информ-объект * (Вид, concept)(Содержание1, content)(r_1 , u_1)…(r_n , u_n) ,

где concept – обозначение понятия “протокол переговоров“ или “контракт“, content – К-представление документа, r_1 , ..., r_n - обозначения внешних характеристик документа (задающих метаданные, например, характеристики Авторы, Дата, Язык), u_1, ..., u_n – цепочки, интерпретируемые как значения характеристик документа.

Дополнительные полезные свойства СК-языков с точки зрения построения СП контрактов и протоколов переговоров определяются возможностями явного указания тематических ролей (концептуальных падежей) в структуре семантического представления ЕЯ-текста, отображения содержания фраз с прямой и косвенной речью, со словами “понятие“, “термин”, рассмотрения функций, значениями и/или аргументами которых могут быть множества объектов (Поставщики, Ассортимент, Директор, Персонал и т.д.).

Автором данной работы был проведен сравнительный анализ выразительных возможностей СК-языков и явлений ЕЯ, отражающихся в структуре деловых контрактов и протоколов коммерческих переговоров (Fomichov 1999a, 2002b). Ряд результатов этого анализа был изложен выше. Проведенный анализ позволяет высказать предположение о том., что выразительных возможностей СК-языков достаточно для построения с их помощью формальных представлений контрактов и протоколов коммерческих переговоров.

С другой стороны, выразительные возможности других известных подходов к формальному представлению содержания ЕЯ-текстов недостаточны для построения СП произвольных контрактов и протоколов переговоров. В частности, это относится к теории представления дискурсов, теории концептуальных графов, эпизодической логике, компьютерной семантике русского языка.

Таким образом, аппарат СК-языков открывает новые возможности построения формальных представлений контрактов и протоколов коммерческих переговоров, осуществляемых компьютерными интеллектуальными агентами.

В то же время из проведенного анализа следует, что СК-языки предоставляют уникальный спектр возможностей для представления результатов семантико-синтаксической обработки лингвистическими процессорами дискурсов, являющихся контрактами или протоколами коммерческих переговоров.

5.3. Разработка семантического сетевого языка нового поколения

В параграфе 1.1 отмечалось возникновение во второй половине 1990-х годов нового направления в разработке семантических языков-посредников и ЛП, использующих такие языки. Это направление появилось как следствие разработки японскими учеными Х. Учидой и М. Жу формального языка для представления содержания предложений, названного ими универсальным сетевым языком (UNL, the Universal Networking Language). Первым центральным мотивом для создания языка UNL было стремление устранить языковой барьер между пользователями сети Интернет из разных стран мира. Второй центральный мотив заключался в попытке создать языковые средства, позволяющие представить в едином формате самые разные знания, накопленные человечеством, и, как следствие, создать объективные предпосылки для совместного использования этих знаний разнообразными компьютерными системами по всему миру.

С конца 1990-х годов Институтом передовых исследований ООН Токийского университета координируется ряд проектов в разных странах, цель которых заключается в создании семейства ЛП, преобразующих предложения на различных естественных языках в выражения языка UNL, а также строящих выражения языка UNL по предложениям на разных естественных языках. В целом эта система проектов охватывает 16 естественных языков, включая 6 официальных языков ООН (Uchida, Zhu, Della Senta 1999; Uchida, Zhu 2001).

Научные результаты, полученные в ходе реализации этих проектов, стали предметом обсуждения на международной конференции по универсальным знанию

и языку, состоявшейся в ноябре 2002 г. в Индии. Значительное внимание на этой конференции было уделено различным аспектам представления знаний о мире в едином формате с помощью языка UNL (Zhu, Uchida 2002).

Из результатов глав 3, 4 следует, что выразительные возможности класса СК-языков значительно превышают выразительные возможности языка UNL. В первую очередь, следует отметить, что язык UNL ориентирован на представление содержания только отдельных предложений, но не произвольных связных текстов. Кроме того, весьма ограничены возможности использования языка UNL для представления знаний о мире. Таким образом, по своим выразительным возможностям язык UNL не полностью, а лишь частично соответствует своему названию “универсальный сетевой язык”. Поэтому представляется обоснованной интерпретация языка UNL как одной из возможных версий семантического языка для сети Интернет, или семантического сетевого языка.

В этой связи можно провести аналогию между исследованиями по разработке семантического сетевого языка, одним из вариантов которого является язык UNL, и исследованиями по разработке языков формирования Web-документов. На протяжении 1990-х годов проходил бурный рост сети World Wide Web (Всемирной Паутины), причем для представления информации использовался преимущественно язык разметки гипертекстов HTML. Однако язык HTML не был предназначен для выделения смысловых частей электронных документов, что привело к большим трудностям принципиального характера при поиске документов, удовлетворяющих запросу пользователя.

Поэтому в конце 1990-х годов консорциумом Всемирной Паутины (обычно обозначается сокращением W3C) была начата подготовка к переходу к новым, семантически-структурированным средствам представления информации в Web-документах. Несколько лет предварительного этапа исследований позволили разработать язык для описания метаданных об информационных ресурсах RDF (Resource Description Framework) и язык RDF SSL - систему спецификации схем, являющихся выражениями языка RDF (RDF 1999; RDF SSL 2000), а затем объявить

о развертывании широкомасштабного проекта Семантической Всемирной Паутины (Semantic Web 2001).

Учитывая сказанное, можно предположить, что широко рекламируемый сегодня язык UNL является не окончательной, а лишь начальной версией семантического сетевого языка. Потребности формального представления содержания сложных связных текстов (например, относящихся к медицине, биологии, экономике, технике, экологии, юриспруденции), а также смысловой обработки семантических представлений (СП) таких текстов в рамках базы знаний о мире должны привести в ближайшие годы к разработке семантического сетевого языка нового поколения.

На основании анализа научной литературы по формализации семантики ЕЯ-текстов можно сделать вывод о том, что в качестве такого языка наиболее целесообразно выбрать СК-язык в некотором концептуальном базисе, построенном с учетом опыта разработки языка UNL. Например, в работе (Uchida, Zhu 2001) рассматривается бинарное отношение *ins* (Инструмент) , связывающее обозначение события и обозначение предмета, с помощью которого реализовано событие. Поэтому можно рассматривать концептуальные базисы с бинарным реляционным символом *ins* , имеюшим тип *tp(ins) = {(соб , [об])}* , где *соб* – сорт “событие” , об – базовый тип “объект” .

Проведенный анализ показал, что нетрудно аппроксимировать все выразительнве механизмы языка UNL средствами СК-языков. Главным образом, это обусловлено тем, что правило P[4] предназначено для конструирования формул с именами *n*-арных отношений, и правило P[8] позволяет строить составные обозначения понятий.

Пример 1. Рассмотрим UNL-выражение *to(train(icl > thing), London(icl > city))* ; это выражение, взятое из работы (Uchida, Zhu, Della Senta 1999), обозначает поезд на Лондон. Данное выражение можно аппроксимировать К-цепочкой вида *Назначение (нек поезд * (Конкретизация, вещь), нек город * (Назв, 'Лондон’))* или вида *нек поезд * (Конкретизация, вещь) (Назначение, нек город * (Назв, 'Лондон’))*.

В то же время, аппарат СК-языков предоставляет ряд важных преимуществ по сравнению с UNL с точки зрения разработки семантического сетевого языка нового поколения. Проиллюстрируем несколько таких преимуществ.

Пример 2. Рассмотрим определение Def1= “A flock (английский язык) – это большое количество птиц или млекопитающих (например, овец или коз), собирающихся вместе с определенной целью, такой, как питание, миграция или оборона”. Тогда определение Def1 может иметь следующее К-представление *Expr1*:

*Определение1 (flock, англ-яз, динамич-группа * (Кач-состав,*
*(птица ∨ млекопитающее * (Примеры, (овца ∨ коза)))), S1,*
(Оценка(Колич-элем(S1), большое) ∧ Цель-формирования (S1,
*Нек намерение * (Примеры, (питание ∨миграция ∨оборона))))) .*

Анализ этой формулы позволяет сделать вывод о том, что при построении семантических представлений (СП) ЕЯ-текстов удобно использовать: (1) обозначение 5-арного отношения *Определение1*, (2) составные обозначения понятий (в данном примере использованы выражения *млекопитающее * (Примеры, (овца ∨ коза))* и *динамич-группа * (Кач-состав, (птица ∨ млекопитающее * (Примеры, (овца ∨ коза))))*), (3) имена функций, аргументами и/или значениями которых могут быть множества (в примере использовано имя одноместной функции *Колич-элем* , значением которой является количество элементов множества), (4) составные обозначения намерений, целей (в примере – выражение *нек намерение * (Примеры, (питание ∨миграция ∨оборона))*).

Структура построенного К-представления *Expr1* в значительной мере отражает структуру исходного определения Т1. Между тем, попытка представить содержание этого определения на языке UNL, т.е. с помощью только обозначений бинарных отношений, привела бы к полному разрушению связи между структурой исходного определения Т1 и структурой UNL-представления данного определения.

Пример 3. Пусть D1 – относящийся к биологии и медицине дискурс “Все гранулоциты являются полиморфонуклеарными. Это означает, что их ядра

многодольны”. Тогда дискурсу D1 можно поставить в соответствие следующее К-представление *Expr2*:

(Свойство (произвольн гранулоцит : x1 , полиморфонуклеарный) : P1) ∧
Пояснение (P1, Следует-из (Ситуация (e1, обладание1 (Агент1, x1)*
(Объект1, нек ядро1 : x2)), Свойство (x2, многодольный)))) .

Ключевую роль в построении К-представления *Expr2* сыграло правило P[5], позволившее ввести метку *x1* для обозначения произвольного гранулоцита, метку *x2* для обозначения ядра клетки, и метку *P1* для обозначения семантического представления первого предложения из дискурса D1. Метка *P1* позволяет в структуре СП текста D1 эксплицировать ссылку на смысл первого предложения текста, даваемую сочетанием “Это означает”.

Язык UNL не включает средств представления ссылок на смысл фраз и более крупных фрагментов дискурса. Между тем, последний пример содержит один из наиболее коротких дискурсов такого рода. Учебники в различных областях знаний изобилуют значительно более сложными дискурсами со ссылками на смысл фраз и более крупных фрагментов.

Анализ показывает, что выразительные механизмы языка UNL нетрудно аппроксимировать средствами СК-языков, поскольку правило P[4] позволяет использовать бинарные реляционные символы. В то же время разработка семантического сетевого языка нового поколения.на основе определения класса СК-языков, в частности, позволит: (1) строить не только СП предложений, но и СП сложных связных текстов за счет средств представления ссылок на ранее упомянутые объекты и на смысл фраз и более крупных фрагментов текстов; (2) формировать составные обозначения множеств, понятий, целей интеллектуальных систем и назначений объектов; (3) соединять с помощью логических связок “и” , “или” не только обозначения высказываний, но и обозначения понятий, объектов, множеств объектов; (4) отображать смысловую сруктуру фраз со словами “понятие”, “термин”; (5) рассматривать нетрадиционные функции, аргументами и/или значениями которых могут быть множества объектов, множества понятий, СП текстов, множества СП текстов.

Таким образом, полученные в главах 2 - 4 результаты открывают реальные перспективы разработки семантического сетевого языка нового поколения, выразительные возможности которого будут значительно ближе к выразительным возможностям ЕЯ по сравнению с возможностями языка UNL, предложенного Х. Учидой и М. Жу (Uchida, Zhu, Della Senta 1999; Uchida, Zhu 2001; Zhu, Uchida 2002). Этот вывод опубликован в работе (Fomichov 2004).

5.4. Новые возможности для построения онтологий предметных областей и разработки языков представления знаний

5.4.1. Онтологии и их значение для глобальных информационных сетей

Работа прикладной интеллектуальной системы существенным образом зависит от ее базы знаний. Еще с 1970-х годов развиваются исследования по разработке все более совершенных языков представления знаний (ЯПЗ) в интеллектуальных системах. Важный класс ЯПЗ составляют терминологические языки представления знаний. В отличие от языка логики предикатов, в терминологических ЯПЗ есть специальные единицы, являющиеся обозначениями понятий, и есть средства построения из таких единиц составных обозначений понятий.

Например, во второй половине 1980-х годов и начале 1990-х годов в Германии по заказу фирмы IBM был реализован проект LILOG (LInguistics & LOGic), реализованный институтом представления знаний (Штутгарт) совместно с несколькими университетами. В рамках этого проекта были разработаны новые средства для представления знаний о мире и для проектирования ЕЯ-диалоговых систем (Pletat, von Luck 1990; Pletat 1991; Herzog, Rollinger 1991).

Терминологические языки представления знаний называют также в англоязычной литературе KL-ONE-like languages, потому что первым терминологическим языком представления знаний был язык KL-ONE,

разработанный в конце 1970-х – начале 1980-х (Brachman, Schmolze 1985). Одним из потомков языка KL-ONE стал язык L_{LILOG} , разработанный в проекте LILOG.

Развитие исследований по разработке терминологических ЯПЗ привело в 1990-е годы к появлению нового значения понятия “онтология“. Согласно Большому энциклопедическому словарю под редакцией А.М. Прохорова, изданному в 2000-м году, в философии онтологией называется учение о бытии (в отличие от гносеологии – учения о познании), в котором исследуются всеобщие основы, принципы бытия, его структура и закономерности.

В работах по информатике онтология понимается как спецификация (т.е. описание) концептуализации (Gruber 1993; Guarino 1998; FIPA 1998b). Термин “концептуализация” используется для указания способа, которым интеллектуальная система структурирует знания о мире, восприятие мира. Спецификация концептуализации дает значения терминам из словаря, используемого интеллектуальной системой для обработки знаний и взаимодействия с другими интеллектуальными системами.

На протяжении последнего десятилетия можно было наблюдать постоянный рост интереса исследователей к построению и изучению онтологий. Причина этого заключается в том, что ученые и разработчики компьютерных систем стали заинтересованы в повторном использовании или/и разделении (совместном использовании) знаний системами. Например, в нашей стране опубликованы, в частности, работы (Гаврилова, Хорошевский 2000; Гаврилова, 2001; Нариньяни 2001, 2002) и обстоятельный обзор (Смирнов, Пашкин, Шилов, Левашова 2002а, 2002б), посвященные созданию и применению онтологий.

Созданию и использованию онтологий для разработки системы автоматизированного контроля смысловой полноты технической документации, описывающей поведение оператора летного экипажа и бортовой аппаратуры в различных полетных режимах, посвящены работы (Добров, Лукашевич и др. 2004; Лукашевич 2004).

В основе исследования лежит частичный семантико-синтаксический анализ документации технических систем. Сущность анализа заключается в выделении

понятий, называемых в тексте или ассоциированных с упоминаемыми в тексте объектами и действиями. Выделенные понятия сопоставляются с фреймовой моделью предметной области, отражающей иерархическую сеть понятий и часть их взаимосвязей. Такая модель названа АвиаОнтологией.

Компьютерные системы используют различные понятия для описания предметных областей. Эти различия создают трудности для применения знаний одной системы в другой системе. Предположим, что мы построим онтологии, которые могут служить основой разработки баз знаний многих систем. В этом случае различные системы смогут применять общую терминологию, а это облегчит разделение и неоднократное использование знаний.

Примерно с начала 1990-х годов исследователями многих стран ведется поиск эффективных формальных подходов к построению онтологий и средств программной реализации онтологий. В 1990-е годы и начале 2000-х годов наибольшую известность получили компьютерные онтологии CYC (Lenat 1995; CYC 2001), LOOM (Loom 2001), OIL (Fensel и др., 2000; Horrocks 2000), DAML (DAML 2001).

В 1990-е годы исследования, направленные на создание терминологических ЯПЗ и применение их к построению онтологий, привели к возникновению нового научного направления в области математической теории прикладных интеллектуальных систем – дескриптивной логики. Общим для различных вариантов логик, разработанных в рамках данного направления, является то, что важный подкласс рассматриваемых правильно построенных формул образуют простые и составные обозначения понятий.

В связи с развертыванием широкомасштабного проекта Семантической Всемирной Паутины (Semantic Web) роль онтологий еще более возросла. Исследования по терминологическим ЯПЗ и дескриптивным логикам, опыт разработки в 1990-е годы компьютерных онтологий, а также разработка (в рамках подготовки проекта Семантической Паутины) языка описания метаданных об информационных ресурсах RDF (Resource Description Framework) позволили специалистам Западной Европы и США создать языковую систему для построения

онтологий DAML + OIL (Horrocks, van Harmelen, Patel-Schneider 2001). Эта система, получившая название языка для разметки онтологий (the ontological markup language), сегодня рассматривается специалистами как важная часть теоретического фундамента Семантической Паутины. Язык DAML + OIL основывается на идеях логики предикатов первого порядка и идеях, реализованных в фреймоподобных ЯПЗ.

Несмотря на интенсивность исследования проблемы, выразительные возможности разработанных формальных языков для построения онтологий являются довольно ограниченными. В частности, это относится к построению семантических представлений (СП) определений понятий, в которых либо упоминаются множества или назначения объектов или цели интеллектуальных систем, либо содержатся ссылки на смысл фраз и более крупных частей дискурса.

5.4.2. Анализ возможностей представления знаний о предметных областях средствами СК-языков

Авторы многих публикаций по онтологиям отмечают, что перспективный путь автоматизации конструирования онтологий заключается в разработке и использовании лингвистических процессоров для извлечения знаний из накопленных во всех областях текстов на естественном языке – монографий, статей, научных и технических отчетов, юридических документов и т.д.

Поэтому необходимы значительно более мощные (по сравнению с имеющимися) формальные средства для построения СП (а) ЕЯ-определений понятий и (б) предложений и дискурсов на ЕЯ, выражающих знания о предметной области.

В этой связи представляется целесообразным указать некоторые наиболее важные выразительные возможности СК-языков с точки зрения построения СП определений понятий и формального отображения знаний о предметных областях.

Сначала покажем (примеры 1 – 3), что СК-языки позволяют моделировать выразительные механизмы основных языков дескриптивной логики (другими словами, терминологических ЯПЗ).

Пример 1. Пусть Т1 = “Тинейджер – это человек в возрасте от 13 до 19 лет.” . Тогда на языке L_{LILOG} содержание текста Т1 может быть представлено выражением

teenager = person ∩ with-feature age in [13..19] .

Первым возможным К-представлением (КП) текста Т1 является выражение

*((тинейджер ≡ человек * (Возраст, x)) ∧ ¬ Меньше (x, <13, год>)*

∧ ¬ Больше (x, <19, год>)).

Вторым возможным КП текста Т1 является формула

*((тинейджер ≡ человек * (Возраст, x)) ∧ Диапазон (x, год, 13, 19))* .

Третье возможное КП текста Т1: *Определение (тинейджер, x, (Явл (x, тинейджер) ≡ (Явл (x, человек) ∧ Диапазон (Возраст (x), год, 13, 19))))* .

Пример 2. В языке L_{LILOG} семантическое представление текста Т2 = “Порше 911 – это автомобиль с двумя дверями типа кабрио” может выглядеть следующим образом (Pletat 1991) : *Constant Porsche-911: and (car, doors : {2}, body: {cabrio})*

Возможным К-представлением текста Т2 является выражение

*(Porsche-911 ≡ car * (Doors-number, 2) (Body-type, cabrio))* .

Пример 3. На языке L_{LILOG} информация о том, что различают типы корпуса автомобилей cabrio, coupe, hatch-back, sedan , формально представляется выражением

Sort body-type;

Atoms cabrio, coupe, hatchback, sedan. .

Эту же информацию можно представить с помощью следующего выражения некоторого СК-языка:

Kinds (body_type, (cabrio ∨ coupe ∨ hatch-back ∨ sedan)) .

Мы видим, что логические связки “и”, “или” позволяют соединять не только семантические представления высказываний, но и обозначения понятий.

Анализ показывает, что выразительные возможности языка L_{LILOG} и других разработанных терминологических ЯПЗ являются довольно ограниченными. Это касается построения составных обозначений понятий и целей интеллектуальных систем, описания множеств, отображения содержания ЕЯ-текстов со ссылками на смысл фраз и более крупных частей дискурса. В связи с этим рассмотрим некоторые возможности использования СК-языков в подобных случаях.

Пример 4. В предыдущем параграфе рассматривалось определение Def1 = “A flock (английский язык) – это большое количество птиц или млекопитающих (например, овец или коз), собирающихся вместе с определенной целью, такой, как питание, миграция или оборона” и была построена цепочка *Expr1* , являющаяся СП этого определения.

Определение Def1 взято из определенной книги, опубликованной в определенном году определенным издательством. СК-языки позволяют строить СП определений и других фрагментов знаний в объектно-ориентированной форме, отражая их внешние связи. Например, объектно-ориентированное СП определения Def1 может являться выражением

*нек информ-объект * (Вид, определ)(Содержание1, Expr1)*
*(Источник1, нек словарь * (Название, ‘Longman Dictionary*
of Scientifc Usage’)(Издательство, (Longman-Group-Limited/Harlow ∧
Russky-Yazyk-Publishers/Moscow))(Город, Москва)(Год, 1989)) .

Пример 5. Пусть ТЗ — определение «Евстахиева труба – это канал, ведущий от среднего уха к глотке ». ТЗ можно поставить в соответствие, в частности, следующую К-цепочку, интерпретируемую как СП текста ТЗ:

Определение1 (евстахиева-труба, русск-яз, канал2, x1,
∃z (чел) Вести1 (x1, нек среднее-ухо ∗ (Часть, z), нек глотка ∗ (Часть, z))) .

Пример 6. Пусть Т4 = «Сфигмоманометр — прибор, предназначенный для измерения кровяного давления», тогда Т4 может иметь следующее КП:

(сфигмоманометр ≡ прибор ∗ (Назначение, измерение1 ∗
(Парам, кровяное-давление)(Субъект, произв чел))) .

Семантическая единица *Назначение* в этом КП обозначает бинарное отношение. Если пара (A, B) принадлежит этому отношению, то A является физическим объектом, а B - формальным семантическим аналогом выражения, описывающего назначение этого физического объекта.

Пример 7. Пусть T5 — определение «Тромбин — это фермент, который помогает преобразовать фибриноген в фибрин во время коагуляции». Тогда следующая К-цепочка является возможным КП T5:

*(тромбин ≡ фермент ∗ (Назначение, оказание-помощи * (Действие, преобразование1* (Исх-объект, нек фибриноген)(Результат1, нек фибрин)(Процесс, нек коагуляция)))).*

Примеры, рассматриваемые ниже, покажут выразительные возможности стандартных К-языков в отношении описания семантической структуры дискурсов.

Пример 8. Рассмотрим текст T6 = «Адениновая основа на одной нити ДНК связана только с тиминовой основой противоположной нити ДНК. Подобным же образом, цитозиновая основа связана только с гуаниновой основой противоположной нити ДНК».

Для построения КП T6 полезно следующее пояснение. Молекула дезоксирибонуклеиновой кислоты (молекула ДНК) содержит тысячи нуклеотидов (комбинаций из трех основных элементов: дезоксирибозы, фосфатов и основы). Существует четыре вида основ: аденин, гуанин, цитозин и тимин. Нуклеотиды ДНК-молекулы образуют цепочку, которая формирует две длинные нити, сплетенные друг с другом. Приняв во внимание это замечание, с первым предложением из T6 можно связать КП A1 вида

$$\forall x1\ (\text{днк-молекула})\ (\text{Связывать1 (произв основа1} * (\text{Явл, аденин})\ (\text{Часть, произв нить1} * (\text{Часть, x1}) : y1) : z1, \text{нек основа1} * (\text{Явл, тимин})\ (\text{Часть, нек нить1} * (\text{Часть, x1})\ (\text{Противоположн, y1}) : y2) : z2) \wedge \neg \exists z3\ (\text{основа1})\ (\text{Явл}\ (z3, \text{тимин}) \wedge \text{Часть}\ (z3, y2) \wedge \text{Связывать}\ (z1, z3)) : P1 \quad (5.4.1)$$

В строке *A1* вида (5.4.1) переменные y1 и y2 используются как метки описаний двух нитей произвольной молекулы ДНК x1; переменные z1, z2, z3 помечают основы. Переменная *P1* (имеет сорт «смысл сообщения») используется для обозначения семантического представления первого предложения из Т6. Это позволяет построить компактное СП второго предложения Т6, так как вхождение выражения «подобным же образом» во второе предложение из Т6 означает ссылку на смысл первого предложения. В частности, второе предложение Т6 в контексте первого предложения может иметь К-представление *A2* вида

*(Подобно (P1, P2) ∧ (P2 ≡ ∀x1 (днк-молекула) (Связывать (произв основа1 * (Явл, цитозин) (Часть, произв нить1 * (Часть, x1) : y3) : z4, нек основа1 * (Явл, гуанин) (Часть, нек нить1 * (Часть, x1)(Противоположн, y3) : y4) : z5) ∧ ¬ ∃z6 (основа1) (Явл (z6, гуанин) ∧ Часть (z6, y4) ∧ Связывать (z4, z6)))))* . (5.4.2)

Таким образом, с текстом Т6 можно связать К-цепочку *A3* вида *(A1 ∧ A2)*, где *A1* и *A2* — цепочки видов (5.4.1) и (5.4.2) соответственно. Полученную строку можно рассматривать как возможное КП для текста Т6.

Цепочка *A3* иллюстрирует важную возможность, предоставляемую стандартными К-языками: можно помечать переменными фрагменты цепочек, являющиеся семантическими представлениями сообщений, неопределенных форм глаголов или вопросов. Эта возможность позволяет эффективно описывать структурированные значения дискурсов со ссылками на смысл фрагментов, являющихся сообщениями, целями (советами, пожеланиями) или вопросами. На наличие подобных ссылок в дискурсах часто указывают слова и словосочетания: «эта рекомендация», «например», «то есть», «рассмотренная идея», «другими словами» и ряд других.

Построенное КП *A3* для Т6 иллюстрирует еще одну особенность СК-языков: символ «≡» соединяет переменную *P2* и семантическое представление предложения.

Пример 9. Пусть Т7 = «Термин «цитозин» используется в генетике». Структурированное значение Т7 может быть представлено в виде К-цепочки

*Используется(нек понятие * (Название1, «цитозин»), генетика) .*

Следовательно, СК-языки позволяют описывать структурированные значения, предложений со словами «понятие», «термин» и т. п.

Таким образом, аппарат СК-языков открывает возможности построения новых терминологических языков представления знаний с очень большой выразительной силой и, как следствие, возможности разработки онтологий в произвольных предметных областях, поскольку дает мощные средства построения составных обозначений объектов, множеств, понятий, целей интеллектуальных систем, а также средства отображения ссылок на ранее упомянутые сущности и на смысл предшествующих фраз и более крупных частей связного текста. Перечисленные особенности СК-языков являются основными преимуществами предложенного подхода к формализации предметных областей по сравнению с языком разработки онтологий DAML + OIL, использующимся в проекте Семантической Паутины.

Поэтому одним из возможных применений аппарата СК-языков является совершенствование средств построения онтологий в проекте Семантической Паутины.

5.4.3. Разработка новых языков представления знаний для решения информационно-сложных задач

Анализ научной литературы говорит о том, что существует глубокая связь между проблемой формального описания содержания ЕЯ-текстов и проблематикой разработки информационных технологий (ИТ), основанных на представлении и обработке сложноструктурированных знаний. Как подчеркивается в работе (Кузин 2004), до последнего времени разработчики ИТ для автоматизации решения разнообразных практических задач основное внимание уделяли поиску методов решения алгоритмически-сложных задач. При этом не было широко осознано существование класса информационно-сложных задач, для которых необходимы языки представления знаний о проблемной среде с большими выразительными возможностями. Такие языки должны, в частности, позволять отображать большое

количество различных смысловых аспектов проблемной среды, обрабатывать информацию с разных точек зрения, строить многоуровневые обобщения и интегрировать информацию.

Анализ публикаций Е.С. Кузина по технологии функционально-ориентированного проектирования (ФОП-технологии) программных систем (Кузин 1996 - 2004) показывает, что основные идеи этих публикаций тесно взаимосвязаны **с** понятием онтология. Несмотря на то, что термин онтология не используется, по-существу, в том же смысле применяется термин модель проблемной среды, обозначающий целостную систему взаимосвязанных знаний о проблемной среде.

В качестве одного из ключевых направлений исследования проблематики автоматизации решения информационно-сложных задач указывается создание адекватной теории отображения объективного мира в программной системе (ПС) и разработка на этой основе языков представления знаний (ЯПЗ) о проблемной среде с очень высокими выразительными возможностями.

Список требований к таким ЯПЗ включает: (1) возможности отображения очень большого числа различных смысловых аспектов проблемной среды, которые являются существенными для решения задачи и, следовательно, должны конструктивно учитываться в ПС; (2) наличие средств представления не только детализированной информации о проблемной среде, но и более крупных информационных образований, которые получаются путем многоуровневых обобщений и интеграции информации и позволяют анализировать информацию под разными углами зрения.

В работах (Кузин 2003, 2004) описываются основные черты разработанного языка описания декларативных знаний (ЯОДЗ), удовлетворяющего перечисленным требованиям. Этот язык был создан в рамках новой семантической теории, названной конструктивной семантикой. Разработка ЯОДЗ была использована для создания системы управления базой знаний (СУБЗ), реализованной с помощью инструментальных языков C++ и Java в операционной системе Windows.

Построенная СУБЗ нашла успешное применение в опытно-конструкторских разработках, на ее основе созданы: (а) автоматизированная система “Персонал”,

предназначенная для ведения индивидуализированной информации о личностях, организациях и других объектах и настраиваемая на конкретные применения; (б) системы информационной поддержки управления поставками сложных компьютерных изделий и их сопровождения в течение жизненного цикла (Кузин 2003).

Сопоставление выразительных возможностей СК-языков и основных черт ЯОДЗ, описанных в (Кузин 2003), позволяет говорить о том, что СК-языки обладают всеми ценными свойствами ЯОДЗ. В частности, СК-языки позволяют: (а) формально различать конкретные объекты и понятия (типы почти всех понятий начинаются с вертикальной стрелки); (б) задавать семантические ограничения на аргументы отношений (для этого используется отображение *tp* , являющееся компонентом концептуально-объектной системы); (в) строить составные обозначения множеств объектов, включающие информационную единицу *все* и обозначения чисел, указывающих количество элементов множества.

В то же время есть ряд важных выразительных механизмов, реализованных в СК-языках, которые, насколько можно судить по работам (Кузин 2003, 2004), отсутствуют в ЯОДЗ. В частности, СК-языки позволяют: (а) строить составные обозначения целей интеллектуальных систем и назначений вещей, составные обозначения понятий и более сложные по сравнению с ЯОДЗ составные обозначения множеств, (б) представлять содержание фраз со словами “понятие”, “термин” и содержание дискурсов со ссылками на смысл фраз и более крупных фрагментов дискурса, (в) строить СП определений и других фрагментов знаний в объектно-ориентированной форме, отражая их внешние связи (см. выше пример 4).

Поэтому можно предположить, что научные результаты, полученные в главах 2 - 4 и в данной главе, окажут позитивное влияние на исследования в области автоматизации решения информационно-сложных задач в качестве теоретической базы для разработки ЯПЗ с большими выразительными возможностями, близкими к возможностям ЕЯ.

5.5. Возможности использования СК-языков в проектировании интеллектуальных информационно-поисковых и вопросо-ответных Интернет-систем нового поколения

5.5.1. Актуальность разработки вопросо-ответных Интернет-систем

Хотя сегодня системы информационного поиска в сети Интернет используются сотнями тысяч и миллионами людей во всем мире (в зависимости от языка запросов), их эффективность еще далеко не соответствует пожеланиям, по-видимому, большей части пользователей. Поэтому актуальной остается проблема совершенствования информационного поиска в сети Интернет.

Важный аспект проблемы заключается в том. что современные поисковые Интернет-системы обрабатывают только тематические запросы. Ответом на такой запрос обычно являются ссылки на большое количество документов – от десятков до десятков тысяч. Между тем, конечным пользователям очень часто нужно получить ответ на вопрос, и такой ответ должен быть числовым значением параметра (например, датой, номером телефона), коротким фрагментом текста или несколькими короткими фрагментами. Примерами таких вопросов являются “Сколько может стоить двухместный номер в трехзвездочном отеле в Будапеште?” и “Сколько провинций в Канаде?”.

В связи с этим как в отечественной научной литературе (см., например, Харин 2002), так и в зарубежных публикациях ставится проблема разработки информационно-поисковых Интернет-систем нового поколения, способных не только осуществлять тематический поиск документов (причем более точно и полно по сравнению с существующими системами), но и отвечать на вопросы конечных пользователей. Решение этой задачи является одной из основных целей реализации широкомасштабного проекта Семантической Всемирной Паутины (Semantic Web 2001).

Выразительные возможности класса СК-языков, исследовавшиеся в главе 4 и в предыдущих параграфах данной главы, показывают, что сегодня аппарат СК-

языков является наиболее удобным инструментом с точки зрения решения следующих задач проектирования информационно-поисковых Интернет-систем нового поколения, способных не только осуществлять тематический поиск документов, но и отвечать на вопросы конечного пользователя: (а) построения семантического представления (СП) запроса пользователя, (б) построения СП фрагмента анализируемого ЕЯ-текста (длина фрагмента может быть сколь угодно большой).

С одной стороны, выше было показано, что класс СК-языков обладает наибольшими выразительными возможностями по сравнению с другими известными подходами к формальному представлению содержания ЕЯ-текстов. С другой стороны, в главе 4 была высказана гипотеза о том, что СК-языки удобны для построения СП произвольных текстов деловой прозы.

Проиллюстрируем часть важных выразительных возможностей СК-языков на примерах, относящихся к двум актуальным научно-техническим проблемам. Первая проблема заключается в разработке методов и базирующихся на них компьютерных систем, обеспечивающих общественности в нащей стране доступ к государственным информационным ресурсам. Второй проблемой является разработка юридических полнотекстовых баз данных.

5.5.2. Электронные библиотеки и проблема обеспечения доступа общественности к государственным информационным ресурсам

Развитие гражданского общества в нашей стране существенно зависит от степени доступности государственных информационных ресурсов. Огромную роль в обеспечении доступа общественности к государственным информационным ресурсам должны сыграть электронные библиотеки (Елепов, Марчук, Бобров, Константинов 1997; Когаловский 2000; Калиниченко, Скворцов и др. 2000; Когаловский, Новиков 2000; Марчук, Осипов 2000; Антопольский 2002; Антопольский, Майорович, Чугунов 2005). Электронным библиотекам (ЭлБ) отводится важная роль в федеральной целевой программе “Электронная Россия

(2002 – 2010 годы)". Для обеспечения подлинной широты доступа пользователей ЭлБ к информационным ресурсам необходимы естественно-языковые интерфейсы (ЕЯ-интерфейсы), образующие важный подкласс лингвистических процессоров (ЛП). Кроме того, необходимы ЛП, способные преобразовать текстовый документ на естественном языке (ЕЯ) или фрагмент текстового документа в формальную структуру, отражающую его содержание, или смысл (семантическое представление документа или его фрагмента), а затем сравнить содержание запроса пользователя с этим семантическим представлением (СП).

Одной из первоочередных теоретических задач, связанных с разработкой ЛП указанных видов, является создание эффективных методов формального описания содержания (или смысла, или смысловой структуры) произвольных или почти произвольных текстов деловой прозы на русском и английском языках. Широкий спектр новых возможностей в этом направлении предоставляют СК-языки. Рассмотрим только один пример, далеко не исчерпывающий все такие возможности.

Пример. Пусть Т1 = "Какие решения правительства за 2000 – 2004 годы направлены на улучшение завоза продовольствия или расширение строительства жилья на северном побережье Восточной Сибири?". Тогда СП запроса Т1 может являться следующим выражением Е1 некоторого СК-языка:

*Вопрос (S1, (Качеств-состав(S1, решение1) ∧ Описание (произвольн решение1 * (Элемент, S1) : x1, (Принято (x1, Правительство (Россия), t1) ∧ Год (t1, (2000 ∨ 2001 ∨ 2002 ∨ 2003 ∨ 2004)) ∧ Цель (x1, (улучшение1 * (Процесс1, завоз1 * (Объект1, нек множ * (Качеств-состав, продукт-питания))(Место2, нек побережье * (Относит-расположение, север)(Регион, Восточн-Сибирь) : x2) ∨ расширение1 * (Процесс1, строительство * (Объект1, нек множ * (Качеств-состав, дом 1 * (Вид1, жилой)) : S3)(Место0, x2)))))).*

В построенном выражении Е1 отражены, в частности, следующие особенности предлагаемого подхода к формализации содержания ЕЯ-текстов.

1. Используются информационные единицы *нек* ("некоторый"), *произвольн* ("произвольный").

2. Можно строить составные обозначения: (а) понятий, характеризующих объекты: *дом 1 * (Вид1, жилой)* ; (б) понятий, характеризующих множества объектов: *множ * (Качеств-состав, продукт-питания)* ,
(в) множеств объектов: *нек множ * (Качеств-состав, продукт-питания)* ,
*нек множ * (Качеств-состав, дом 1 * (Вид1, жилой)) : S3*.
3. Можно присоединять (с помощью двоеточия) метки к составным обозначениям объектов. Например, в выражении

*нек побережье * (Относит-расположение, север)(Регион, Восточн-Сибирь) : x2*

переменная *x2* является меткой, поставленной в соответствие северному побережью Восточной Сибири.

4. Можно строить составные обозначения целей:

*улучшение1 * (Процесс1, завоз1 * (Объект1, нек множ * (Качеств-состав, продукт-питания))(Место2, нек побережье * (Относит-расположение, север)(Регион, Восточн-Сибирь) : x2)).*

5. Можно представлять сложные цели, соединяя логическими связками ∧ (и), ∨ (или) обозначения более простых целей:

*(улучшение1 * (Процесс1, завоз1 * (Объект1, нек множ * (Качеств-состав, продукт-питания) : S2)(Место2, нек побережье * (Относит-расположение, север)(Регион, Восточн-Сибирь) : x2)) ∨ расширение1 * (Процесс1, строительство * (Объект1, нек множ * (Качеств-состав, дом 1 * (Вид1, жилой)) : S3)(Место0, x2))).*

6. Логические связки ∧ (и), ∨ (или) могут соединять обозначения объектов (а не только высказываний, как в логике предикатов): *(2000 ∨ 2001 ∨ 2002 ∨ 2003 ∨ 2004).*
7. СК-языки позволяют связать с обозначением множества простое или составное обозначение понятия, являющегося концептуальной характеристикой каждого элемента этого множества:

Качеств-состав(S1, решение1), Качеств-состав(S2, продукт-питания),
*Качеств-состав(S3, дом 1 * (Вид1, жилой)).*

Глава 6
МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ЛИНГВИСТИЧЕСКОЙ БАЗЫ ДАННЫХ

Построена широко применимая формальная модель лингвистической базы данных (ЛБД), т.е. базы данных, которая содержит информацию, используемую алгоритмом семантико-синтаксического анализа текстов для построения по ЕЯ-тексту его семантического представления (СП). Эта модель описывает логическую структуру ЛБД ЕЯ-интерфейсов баз данных и других прикладных компьютерных систем. В предложенной модели выражения стандартных К-языков (СК-языков) используются в качестве семантических единиц, соответствующих лексическим единицам, и в качестве СП естественно-языковых текстов. Предложена новая структура данных (названная матричным семантико-синтаксическим представлением текста), используемая в качестве промежуточной формы отображения семантической структуры входного ЕЯ-текста, для последующего построения СП текста. Разработан предметно-независимый метод преобразования входного ЕЯ-текста (вопроса, команды, сообщения) из подъязыка русского языка в его СП, являющееся выражением некоторого СК-языка.

6.1. Постановка задачи

Огромный рост парка персональных компьютеров, разработка многочисленных баз данных (БД) и баз знаний (БЗ) привели к тому, что к этим БД и БЗ получил доступ широкий круг пользователей, не являющихся программистами и не изучавших какие-либо формальные языки. Тем не менее, у этих пользователей возникает необходимость взаимодействия с такими БД или БЗ для решения профессиональных задач либо в повседневной жизни. Стремительное развитие сети Интернет существенно усиливает эту тенденцию: у очень большого

количества людей появляется желание получить какую-то информацию из источников, удалённых от терминала пользователя на тысячи километров.

Все эти факторы способствовали во второй половине 1990-х годов – первой половине 2000-х годов значительному усилению внимания к разработке и применению естественно-языковых интерфейсов (ЕЯ-интерфейсов) БД и БЗ. Одним из свидетельств возникновения новой ситуации в области проектирования и применения ЕЯ-интерфейсов является разработка научно-исследовательским центром фирмы Microsoft® англоязычного интерфейса баз данных English Query. С конца 1990-х годов этот интерфейс поставляется вместе с сервером SQL 6.5 или SQL 7.0 и может встраиваться в состав Web-узлов пользователей. Интерфейс English Query позволяет задавать вопросы к реляционной БД на структурно-ограниченном английском языке. Предусмотрены средства адаптации интерфейса к новым предметным областям (English Query 2000; Snyder 2001)..

Учитывая сказанное выше, можно сделать вывод о том, что насыщенность персональными компьютерами крупнейших промышленных и научных центров нашей страны, бурное развитие сети Интернет в последние годы говорят о большой актуальности проблемы создания интеллектуальных интерфейсов, предоставляющих массовому пользователю возможность взаимодействия с базами данных и/или базами знаний на структурно-ограниченном русском языке (РЯ) (Попов 1999).

Исследования по разработке ЛП для РЯ и некоторых других языков (в первую очередь, английского и французского) развиваются в нашей стране в течение более чем трёх десятилетий. В конце 1990-х – начале 2000-х годов потребности практики в развитии лингвистических информационных технологий (ЛИТ) привели к появлению ряда новых интересных проектов ЛП. Такие проекты были реализованы, в частности, в Институте проблем информатики РАН (Кузнецов и др. 2000; Кузнецов, Мацкевич 2001, 2003; Kuznetsov, Matskevich 2002), Институте проблем передачи информации РАН (Богуславский, Иомдин и др. 2000), распределенным коллективом разработчиков интеллектуальной метапоисковой системы “Сириус” из Института программных систем РАН, Институтf системного

анализа РАН и Российского государственного университета дружбы народов (Куршев, Осипов и др. 2002; Осипов, Куршев и др. 2003; Завьялова 2004; Тихомиров, Осипов и др. 2005), в Институте прикладной математики им. М.В. Келдыша (Кулагина 1998, 2001; Агранат, Кулагина 2000), РосНИИ информационных технологий и систем автоматизированного проектирования (Курбатов, Попов 2001; Курбатов 2002), РосНИИ искусственного интеллекта (Жигалов, Соколова 2001; Жигалов 2002; Жигалов В.А., Жигалов Д.В. 2002), ВИНИТИ (Кузнецов, Солнцева, Деревянкин, Закамская 2001), на факультете вычислительной математики и кибернетики МГУ им. М.В. Ломоносова (Мальковский, Шикин 1998; Болдасов, Соколова 2002), в МВТУ им. Н.Э. Баумана (Смирнов, Андреев, Березкин, Брик 1997), фирмой Abbey Software House (Перцова, Перцов 2002), ООО “Гарант-Парк-Интернет” (Ермаков 2002; Киселев, Ермаков, Плешко 2004), в Санкт-Петербургском государственном университете и Санкт-Петербургском экономико-математическом институте РАН (Тузов 2001; Каневский, Тузов 2002; Лезин, Тузов 2003), в Институте высокопроизводительных вычислений и баз данных Санкт-Петербургского государственного технического университета (Писарев, Самсонова 2002).

Важное значение для развития в нашей стране ЛИТ имеет также реализация нескольких проектов компьютерных семантических словарей (Лахути, Рубашкин 1998 - 2000; Леонтьева 2000 - 2002; Леонтьева, Семенова 2001 – 2003; Мальковский, Соловьев 2002 – 2004; Пацкин 2004).

Это продвижение в области проектирования и применения ЛП во многом было подготовлено теоретическими результатами и опытом проектирования ЛП, полученными в 1980-х годах и начале 1990-х годов, в частности, Ю.Д. Апресяном, И.М. Богуславским (Апресян, Богуславский и др. 1981, 1989), Г.Г. Белоноговым, И.А. Большаковым, В.М. Брябриным, А.В. Гладким (Гладкий 1985), Б.Ю. Городецким, В.И. Дракиным, А.П. Ершовым, Е.С. Кузиным, И.П. Кузнецовым, О.С. Кулагиной (Кулагина 1979, 1996), С.С. Курбатовым, Д.Г. Лахути, В.Ш. Рубашкиным (Рубашкин 1989; Лахути, Рубашкин 1993), Н.Н. Леонтьевой (Леонтьева 1981, 1986), Л.В. Литвинцевой, Ю.Я. Любарским (Любарский 1990),

М.Г. Мальковским (Мальковский 1985), Л.И. Микуличем, А.С. Нариньяни, А.П. Новоселовым, Г.С. Осиповым (Осипов 1990, 1997), Н.В. Перцовым, Р.Г. Пиотровским, Д.А. Поспеловым, Э.В. Поповым, А.Б. Преображенским (Попов 1982, 1987; Попов, Преображенский 1990; Дракин, Попов, Преображенский 1988), Г.В. Рыбиной, Г.В. Сениным, А.М. Степановым, В.А. Тузовым (Тузов 1984), В.С. Файном, Г.К. Хахалиным, В.Ф. Хорошевским, Г.С. Цейтиным, Л.Л. Цинманом (Апресян, Цинман 1982, Цинман 1986), а также рядом других исследователей.

Несмотря на появление в конце 1990-х – начале 2000-х годов новых примеров применения на практике в нашей стране лингвистических процессоров, можно констатировать, что в этот период в целом недостаточно внимания уделялось разработке эффективных формальных средств и методов проектирования ЛП.

Наибольшие трудности при разработке ЛП связаны с выполнением преобразования “ЕЯ-текст ➔ Семантическое представление (СП) текста”, где под СП ЕЯ-текста понимается формальная структура, отражающая содержание (или смысл) ЕЯ-текста. Однако анализ как отечественных, так и зарубежных публикаций показывает, что при разработке преобразователей ЕЯ-текстов в СП текстов крайне недостаточно используются формальные средства. Это выражается в неформальном и фрагментарном описании (а) структуры лингвистической базы данных (ЛБД), т.е. базы данных с морфологической и семантико-синтаксической информацией о лексических единицах, используемой алгоритмом семантико-синтаксического анализа текстов для построения по ЕЯ-тексту его семантического представления (СП) и (б) методов обработки информации основными подсистемами преобразователя “ЕЯ-текст ➔ СП текста”.

Основная часть исследований по разработке ЕЯ-интерфейсов и ЛП других видов была реализована для английского языка, синтаксис которого существенно отличается от синтаксиса русского языка (РЯ). В отличие от английского языка, РЯ относится к классу сильно флективных языков. Это выражается в том, что слова РЯ могут изменяться; например, окончания существительных меняются в зависимости от грамматического падежа и числа, окончания глаголов зависят от времени и лица и т.д. Другой важной особенностью РЯ является весьма свободный

порядок слов; например, в предложениях с глаголом в действительном залоге подлежащее может располагаться как перед сказуемым, так и после сказуемого.

Чрезвычайно существенно то, что полные описания информационного и программного обеспечения англоязычных ЛП, как правило, недоступны специалистам в нашей стране. Кроме того, одним из следствий экономической ситуации, сложившейся в 1990-е годы в нашей стране, является отсутствие даже в центральных библиотеках огромного количества публикаций в области разработки ЛП, опубликованных за рубежом в 1990-е и 2000-е годы на английском и некоторых других языках. Все это серьезно затрудняет подготовку в нашей стране специалистов в области проектирования ЛП и сужает возможности принятия оптимальных проектных решений, приводит к дополнительным трудозатратам на разработку ЛП.

Таким образом, актуальной является проблематика разработки методов формального описания структуры ЛБД, а также таких методов семантико-синтаксического анализа текстов из представляющих практический интерес подъязыков русского языка, которые более широко используют формальные средства описания входных, промежуточных и выходных данных по сравнению с известными методами.

Разработка ЛП многих видов, например, ЕЯ-интерфейсов больших БД, отличается значительной трудоемкостью. В связи с этим в параграфе 1.1. данной книги была выдвинута гипотеза о том, что в долговременной перспективе сокращению затрат и времени на разработку семейства ЛП в рамках одной организации или нескольких взаимодействующих организаций будет способствовать реализация в проектировании информационного и алгоритмического обеспечения ЛП следующих двух принципов:

(1) *принципа стабильности* используемого языка семантических представлений (ЯСП) по отношению к многообразию решаемых задач, многообразию предметных областей и многообразию программных сред (стабильность понимается как использование единой системы правил для построения

конструкций ЯСП и варьируемого набора первичных информационных единиц, определяемого предметной областью и решаемой задачей);

(2) *принципа преемственности* алгоритмического обеспечения ЛП на основе использования одной или нескольких совместимых формальных моделей лингвистической БД и единых формальных средств предствления промежуточных и окончательных результатов семантико-синтаксического анализа ЕЯ-текстов по отношению к многообразию решаемых задач, предметных областей и программных сред (преемственность понимается как многократное использование в различных лингвистических процессорах алгоритмов, реализуемых основными подсистемами ЛП).

Теоретическую основу для реализации принципа стабильности используемого ЯСП создают результаты, изложенные в главах 2 – 4 данной монографии. В главе 3 определен класс стандартных К-языков (СК-языков), позволяющих строить СП ЕЯ-текстов в произвольных предметных областях.

Данная глава базируется на результатах, отраженных в предыдущих главах и направлена на создание значительной части предпосылок для реализации принципа преемственности при проектировании алгоритмического обеспечения лингвистических процессоров..

В данной главе и главе 7 ставится и решается задача разработки нового метода преобразования ЕЯ-текста в семантическое представление для проектирования семантико-синтаксических анализаторов текстов из представляющих практический интерес подъязыков РЯ. С этой целью ставятся и достигаются следующие цели:

1. Формализовать структуру лингвистической базы данных, позволяющей устанавливать возможные смысловые отношения, в частности в сочетаниях «Глагол + Предлог + Существительное», «Глагол + Существительное», «Существительное1 + Предлог + Существительное2», «Число + Существительное», «Прилагательное + Существительное», «Существительное1 + Существительное2», «Причастие + Существительное», «Причастие + Предлог + Существительное», «Вопросительно-относительное местоимение + Глагол», «Предлог + Вопросительно-относительное местоимение + Глагол»,

“Местоименное наречие, играющее роль вопросительного слова + Глагол”, “Глагол + Обозначение числового значения параметра (обозначение числа + обозначение единицы измерения)”.

2. Формализовать структуру данных, используемых в качестве промежуточной формы отображения семантической структуры входного ЕЯ-текста, для последующего построения СП входного текста.
3. На основе решения задач 1 и 2 разработать предметно-независимый метод преобразования входного ЕЯ-текста (вопроса, команды, сообщения) из подъязыка русского языка в его СП.

В данной главе аппарат СК-языков применен к построению широко применимой формальной модели ЛБД. Эта модель описывает логическую структуру ЛБД ЕЯ-интерфейсов интеллектуальных баз данных и других прикладных компьютерных систем. В построенной модели выражения СК-языков используются, во-первых, в качестве семантических единиц, соответствующих лексическим единицам, и, во-вторых, для сборки СП текстов из элементов ЛБД.

Новый метод преобразования ЕЯ-текстов в их семантические представления (СП) предусматривает использование предложенного автором матричного семантико-синтаксического представления входного текста (это понятие было введено в работах (Fomichov 1998, 2002; Фомичев, Волчков 1999) как промежуточного представления при переходе от ЕЯ-текста к СП текста. При этом не используется традиционное синтаксическое представление текста.

Работоспособность предложенного метода доказана разработкой автором сложного структурированного алгоритма семантико-синтаксического анализа текстов из подъязыков естественного (русского) языка и успешным созданием на его основе семейства экспериментальных русскоязычных интерфейсов баз данных и баз знаний, реализованных в программных средах Турбо-Паскаль, версия 7.0, Си, Си++, Delphi 4.0, 5.0, PHP.

6.2. Формализация дополнительных требований к языку построения семантических представлений входных текстов лингвистического процессора

При построении семантических представлений (СП) ЕЯ-текстов в разных предметных областях возникает потребность в использовании небольшого инвариантного набора информационных единиц, в частности, предназначенных для формирования СП вопросов, команд и описаний множеств.

Предположения 1 – 7, сформулированные в процессе исследования выразительных возможностей стандартных К-языков (СК-языков) в главе 4, отражают существо важной части таких потребностей.

В этой и следующей главах мы будем рассматривать концептуальные базисы., для которых выполняются Предположения 1 – 7. Абстрагируясь от математических деталей, можно сказать, что такие концептуальные базисы будут названы размеченными концептуальными базисами. Цель вводимых ниже определений заключается в том, чтобы формально задать понятие размеченного концептуального базиса.

Определение. Пусть *B* – произвольный концептуальный базис, *St*(*B)* – множество сортов базиса *B, P(B)* – выделенный сорт “смысл сообщения”, *X*(*B)* – первичный информационный универсум базиса *B.* Тогда упорядоченный набор *Qmk* вида

(сит,Vsit, Ситуация, Вопрос, лог, ист, ложь, Ист-знач) (6.2.1)

называется разметкой вопросов для концептуального базиса B ⇔ когда

сит, лог ∈ St(B) \ { P(B) } , *X(B)* включает несовпадающие элементы *Ситуация, Вопрос, ист, ложь, Ист-знач,* и выполняются Предположения 1, 4, 6.

Определение. Пусть B – произвольный к.б. Тогда упорядоченный набор *Setmk* вида

(нат, Nt, множ, Колич, Кач-состав, Предм-состав, произв, все, Элем) (6.2.2)

называется теоретико-множественной разметкой базиса B ⇔ *нат ∈ St(B) \ { P(B) },*

Nt - подмножество первичного информационного универсума *X*(*B)*; *множ, Колич, Кач-состав, Предм-состав, произв, все, Элем* – различные элементы множества *X*(*B),* и для компонентов этого набора выполняются Предположения 2, 3, 5.

Определение. Пусть *B* – произвольный к. б., *Qmk* – разметка вопросов вида (6.2.1) для *B,* тогда упорядоченный набор *Cmk* вида

(интс, мом, #сейчас#, #Оператор#, #Исполнитель#, Команда) (6.2.3)

будет называться разметкой команд для базиса *B,* согласованной с разметкой вопросов *Qmk* ⇔ когда *интс, мом, #сейчас#, #Оператор#, #Исполнитель#, Команда* – различные элементы множества X(B); *интс, мом ∈ St(B) \ {P, сит, лог}* и выполняется Предположение 7.

Совокупность формальных понятий, рассмотренных выше в этом подразделе, позволяет сделать заключительный шаг и объединить эти понятия в определении класса размеченных концептуальных базисов.

Определение. *Размеченным концептуальным базисом (р.к.б.)* называется произвольный упорядоченный набор *Cb* вида

(B, Qmk, Setmk, Cmk) , (6.2.4)

где *B* – произвольный концептуальный базис, *Qmk* – разметка вопросов вида (4.2.1) для *B, Setmk* – теоретико-множественная разметка для *B, Cmk* – разметка команд вида (4.2.3) для *B,* согласованная с разметкой вопросов *Qmk,* и выполняются следующие условия: (а) все компоненты наборов *Qmk, Setmk, Cmk* , кроме компонента *Nt* набора *Setmk,* является несовпадающими (различными) элементами первичного информационного универсума X(B);

(б) если *Stadd = { сит, лог, интс, мом, нат },* то *Stadd* –подмножество множества *St(B) \ { P(B) },* причем любые два различные элементы подмножества *Stadd* являются несравнимыми как для отношения общности *Gen,* так и для отношения совместимости *Tol*; (в) если s – произвольный элемент подмножества *Stadd,* то *s* и *P* несравнимы как для отношения *Gen,* так и для и для отношения *Tol*.

Определение. Будем говорить, что размеченный концептуальный базис *Cb* является размеченным базисом стандартного вида ⇔ когда *Cb* – упорядоченный

набор вида (6.2.4), *Qmk* – набор вида (6.2.1), *Setmk* – набор вида (6.2.2) и *Cmk* – набор вида (6.2.3).

В дальнейшем будем рассматривать размеченные концептуальные базисы только стандартного вида.

Класс языков {*Ls(B)* | *B* – первый компонент произвольного р.к.б. *Cb*} будем использовать в качестве семантических языков при рассмотрении соответствий вида "ЕЯ-текст → Семантическое представление текста".

Данный класс языков удобен для построения семантических представлений высказываний, вопросов и команд, причем тексты каждого из указанных видов могут включать составные описания множеств. Многочисленные примеры использования выражений языков этого класса в качестве СП высказываний, вопросов и команд можно найти в главах 3 и 4.

6.3. Текстообразующие системы

Представим формально сведения об элементах, из которых состоят ЕЯ-тексты.

6.3.1. Морфологические базисы

Морфологией называется та часть языкознания, которая изучает закономерности изменения слов и словосочетаний (по числам, падежам, временам и т.д.). Лингвистическая база данных (ЛБД) должна включать *морфологическую базу данных* (МБД), содержание которой зависит от рассматриваемого языка. В отличие от английского языка русский язык является сильно флективным, т.е. слова в нем могут изменяться многими способами. Поэтому, если для английского языка МБД является лостаточно простой, то для русского языка (РЯ) это не так. Формализации морфологии РЯ посвящено много публикаций. Однако для разработки структурированного алгоритма семантико-синтаксического анализа текстов РЯ потребовалось предложить новый, более общий взгляд на морфологию русского языка по сравнению с имеющимися публикациями. Цель заключалась в том, чтобы

указать место морфологического анализа как части семантико-синтаксического анализа ЕЯ-текстов, избегая излишне детального рассмотрения проблем морфологии РЯ. Для достижения этой цели вводятся понятия морфологического детерминанта, морфологического пространства, морфологического базиса и морфологического базиса русскогоязычного типа (Р-типа).

Определение. *Морфологическим детерминантом* (*М-детерминантом*) будем называть произвольную упорядоченную тройку вида

$$(m, n, maxv)\ , \qquad (6.3.1)$$

где m,n - положительные целые числа; *maxv* - отображение из множества $\{1, 2, \ldots m\}$ в множество неотрицательных целых чисел N^+.

Пусть *Det* - М-детерминант вида (6.3.1), тогда m будем интерпретировать как количество всевозможных различных признаков (называемых морфологическими) слов из рассматриваемого языка; n - как максимальное количество различных наборов морфологических признаков, которые могут быть связаны с одним словом. Если $1 \le i \le m$, то *maxv(i)* интерпретируется как максимальное значение признака с номером i (см. рис. 6.1).

Например, со словом "книги" может быть связано три набора значений морфологических признаков (если "книги" - словоформа в единственном числе, то эта словоформа находится в родительном падеже; если "книги" - словоформа во множественном числе, то она может быть как в именительном, так и в винительном падежах). Поэтому $n \ge 3$.

Набор 1	Набор 2	…	Набор n

.

Рис.6.1. Структура массива морфологических признаков, связанных с одной словоформой.

Условимся считать, что морфологические признаки с порядковыми номерами 1 и 2 - это признаки "часть речи" и "подкласс части речи". Поэтому каждое целое k,

такое, что $1 \le k \le maxv(1)$, будем интерпретировать как обозначение какой-то части речи, и каждое целое r, такое, что $1 \le r \le maxv(2)$, будем интерпретировать как обозначение какого-то подкласса некоторой части речи.

Рис. 6.2. иллюстрирует структуру одного набора морфологических признаков с учетом этого соглашения.

Код части речи P_1	Код подкласса ч. речи P_2	Код признака P_3		Код признака P_k	Код признака P_m
1	2	3	…	k	

Рис. 6.2. Структура одного набора значений морфологических признаков.

Будем предполагать, что каждое слово из рассматриваемого языка относится только к одной части речи и к одному подклассу части речи. С одной стороны, это предположение выполняется для чрезвычайно широкого подмножества русского языка и, например, немецкого языка. С другой стороны, такое предположение позволит избежать усложнения (без ущерба для приложения) предлагаемой формальной модели ЛБД.

Определение. Пусть *Det* - М-детерминант вида (6.3.1.). Тогда *морфологическим пространством*, задаваемым детерминантом *Det*, называется множество *Spmorph*, состоящее из всех упорядоченных наборов вида

$$(x_1,\ldots,x_m, x_{m+1},\ldots, x_{2m}, x_{2m+1},\ldots, x_{nm}) \ , \qquad (6.3.2)$$

где: (а) для каждого k=1,…,n-1 $x_{km+1}=x_1$, $x_{km+2}=x_2$; (б) для каждого k=1,…,n и каждого q, такого, что $(k-1)m+1 \le q \le km$, выполняется неравенство $0 \le x_q \le maxv(q-(k-1)m)$.

Условия (а), (б) из данного определения интерпретируются следующим образом. В элементе морфологического пространства вида (6.3.2) x_1 - код части речи, и этот

код расположен во всех позициях, удаленных на расстояние *m, 2m,…, (n-1)m* от позиции 1; x_2 - код подкласса части речи, этот код расположен во всех позициях, удаленных на расстояние *m, 2m,…, (n-1)m* от позиции 2.

Отображение *maxv* для каждого числового кода названия признака q, где $1 \le q \le m$, задает диапазон его значений [1, *maxv(q)*]. Таким образом, для каждой позиции q, где $1 \le q \le m$, $0 \le x_q \le maxv(q)$ для набора вида (6.3.2). Если $1 \le q \le m$, x_q – компонент набора вида (6.3.2) , и $x_q = 0$, то это означает, что словоформа, которой соответствует данный элемент морфологического пространства, не обладает признаком с номером q. Например, у существительных нет признака “время”.

Каждый компонент x_s элемента морфологического пространства вида (6.3.2.), где $s=q+m,q+2m,\ldots,q+(n-1)m$, интерпретируется как какое-то возможное значение морфологического признака, что и в случае элемента x_q. Поэтому неравенство $0 \le x_s \le maxv(s-(k-1)m)$ задает диапазон допустимых значений элемента x_s, где целое число k в пределах от 1 до n однозначно определяется условием $(k-1)m+1 \le s \le km$.

Вводимое ниже определение морфологического базиса дает новую математическую интерпретацию понятия "морфологическая база данных". Если временно абстрагироваться от математических деталей, то под морфологическим базисом мы будем понимать произвольный упорядоченный набор *Morphbs* вида

(Det, A, W, Lecs, lcs, fmorph, propname, valname), (6.3.3)

где Det - морфологический детерминант, а остальные компоненты интерпретируются следующим образом. A – это произвольный алфавит (конечное множество символов); из элементов A образуются словоформы естественного языка. Пусть A^+ - множество всех непустых цепочек в алфавите A. Тогда W - это конечное подмножество A^+, элементы которого рассматриваются как слова и фиксированные словосочетания (например, "в течение"), используемые для построения ЕЯ-текстов. Элементы множества W будем называть словоформами. Lecs - это конечное подмножество множества W, элементы которого называются лексемами и интерпретируются как базовые формы слов и фиксированных словосочетаний (существительное в единственном числе и именительном падеже,

прилагательное в единственном числе, именительном падеже, мужском роде и т.д.).

Компонент *lcs* – это отображение вида $W \rightarrow Lecs$, которое каждой словоформе ставит в соответствие некоторую лексему; *fmorph* – это отображение, которое словоформе *wd* из *W* ставит в соответствие некоторый элемент морфологического пространства *Spmorph(Det).* Компонент *propname* (сокращение от *property-name*) является отображением, которое числовому коду морфологического признака словоформы ставит в соответствие цепочку - его имя. Например, может выполняться соотношение *propname(1)* = *часть-речи.* Точнее, это отображение $propname: \{1,2,\ldots,m\} \rightarrow A+ \setminus W$, где \ - знак теоретико-множественной разности.

Компонент *valname* – это отображение, которое числовому коду морфологического признака k и числовому коду значения данного признака p ставит в соответствие буквенное обозначение данного признака *valname(k, p).* В частности, может выполняться соотношение v*alname(1,1)=глаг* .

Определение. Пусть *A,B* – произвольные непустые множества и $f:A \rightarrow B$ - отображение из A в B. Тогда область значений *Range(f)* – это множество всех таких *y*, что существует такой *x* из *A* , для которого *f(x)= y.*

Определение. Морфологическим базисом называется произвольный упорядоченный набор *Morphbs* вида (6.3.3), где *Det* – М-детерминант вида (6.3.1), *A*-произвольный алфавит, *W* - конечное подмножество множества *A+* (множества всех непустых цепочек в *A*), *Lecs* – конечное подмножество множества *W*, $lcs: W \rightarrow Lecs$ - отображение из *W* на *Lecs* , $fmorph: W \rightarrow Spmorph(Det)$ – отображение из W в морфологическое пространство, порождаемое детерминантом *Det*, *propname* - отображение из $\{1,2,\ldots,m\}$ *в* $A+ \setminus W$, *valname* – частичное отображение из декартового произведения N+ x N+ в множество $A+ \setminus (W \cup Range(propname))$, определенное для пары (i,j) из $N+ \times N+ \Leftrightarrow 1 \leq i \leq m,\ 1 \leq j \leq maxv(i).$

6.3.2. Морфологические базисы Р-типа (русскоязычного типа)

Определение. Морфологический базис вида (6.3.3) называется морфологическим базисом Р-типа (русскоязычного типа) ⇔ выполняются следующие условия:

А – алфавит русского языка, дополненный знаком ‘ -‘ ;

Propname (1) = часть-речи

Propname (2) =подкласс-части-речи

Propname (3) = падеж

Propname (4) = число

Propname (5) = род

Propname (6) = залог

Propname (7) = время

Propname (8) = наклонение

Propname (9) = вид

Propname (10) = лицо

Propname (11) = возвратность

Valname (1,1) = глагол

Valname (1,2) = сущ

Valname (1,3) = прилаг

Valname (1,4) = предлог

Valname (1,5) = местоим

Valname (1,6) = прич

Valname (1,7) = наречие

Valname (1,8) = колич-числит

Valname (1,9) = порядк-числит

Valname (1,10) = союз

Valname (2,1) = сущ-нарицат

Valname (2,2) = сущ-собств

Valname (2,3) = личн-местоим

Valname (2,4) = вопр-относ-местоим

Valname (2,5) = местоим-наречие

Valname (2,6) = глаг-в-изъявит-накл

Valname (2,7) = глаг-в-повелит-накл

Valname (2,8) = глаг-в-неопред-форме

Valname (2,9) = действит-причастие

Valname (2,10) = страдат-причастие

Valname (3,1) = именительн

Valname (3,2) = родительн

Valname (3,3) = дательн

Valname (4,4) = винительн

Valname (5,5) = творительн

Valname (6,6) = предложный

Valname (4,1) = ед. числ.

Valname (4,2) = множ. числ.

Valname (5,1) = жен. род.

Valname (5,2) = муж. род.

Valname (5,3) = сред. род.

Valname (6,1) = действ

Valname (6,2) = страд

Valname (7,1) = прошед-время

Valname (7,2) = наст-время

Valname (7,3) = буд-время

Valname (8,1) = изъявит

Valname (8,2) = повелит

Valname (8,3) = сослаг

Valname (9,1) = несоверш-вид

Valname (9,2) = соверш-вид

Valname (10,1) = 1-ое-лицо

Valname (10,2) = 2-ое-лицо

Valname (10,3) = 3-е-лицо.

Valname (11,1) = действ

Valname (11,2) = страд

Valname (11,1) = вз

Valname (11,2) = нвз .

Здесь *прич* - обозначение части речи “причастие”, *действ* и *страд* – признаки действительного и страдательного залогов глагола, *вз* и *нвз* – признаки возвратных и невозвратных глаголов и причастий.

Определение. Пусть *Morphbs* –морфологический базис вида (6.3.3). Тогда *Parts(Morphbs) = { valname (1,1), … , valname(1,maxv(1)) },*

Subparts(Morphbs) = { valname (2,1), … , valname(2, maxv(2)) }.

Таким образом, *Parts(Morphbs)* – множество названий частей речи, *Subparts(Morphbs)* - множество названий подклассов частей речи для морфологического базиса *Morphbs.* Следовательно, если *Morphbs* – морфологический базис Р-типа, то *Parts(Morphbs)* $\supseteq$ *{ глаг, сущ, прилаг, предлог, местоим, прич, наречие, колич- числит, порядк- числит, союз}, Subparts(Morphbs)* $\supseteq$ *{сущ-нарицат, сущ-собств, вопр-относ-местоим}.*

Пусть *Morphbs* - морфологический базис вида (6.3.3), $z \in Spmorph(Det)$ - произвольный элемент морфологического пространства, и $1 \le i \le mn$, тогда *z [i]* – *i*-й компонент набора z (очевидно, *z* имеет $m \cdot n$ компонентов).

Определение. Пусть *Morphbs* - морфологический базис вида (6.3.3). Тогда отображение *prt* из *W* в *Parts(Morphbs*) и отображение *subprt* из *W* в *Subparts(Morphbs*) задаются следующим образом: для произвольной словоформы $d \in W$ *prt(d) = valname (1, morph(d) [1]), subprt(d) = valname (2, morph(d)[2])* .

Таким образом, *prt(d)* и *subprt(d)* – это соответственно названия части речи и подкласса части речи, к которым относится словоформа d.

Пример. Морфологический базис *Morphbs* Р-типа может быть определён так, что *W э контейнеров, откуда; prt(контейнеров) = сущ, subprt(контейнеров) = сущ-нарицат, prt(откуда) =наречие, subprt(откуда) =местоим-наречие*.

6.3.3. Понятие текстообразующей системы

В текстах могут встречаться не только слова, но и выражения, которые являются числовыми значениями различных признаков, например, 30°, 108%, 90 км/ч, 120 км. Назовём такие выражения *конструктами* и будем считать их единицами текстов. Это означает, что при построении формальной модели лингвистической базы данных, мы будем, например, рассматривать выражение 120_км как символ.

Разрабатывая компьютерные программы, конечно, нужно учитывать, что между “120” и ”км” есть пробел, и “120 км“ - это сочетание из двух элементарных выражений. Но построение всякой формальной модели включает идеализацию сущностей какой-то предметной области, поэтому мы рассматриваем конструкты как символы, т.е. как неделимые выражения.

Кроме слов и конструктов, в текстах встречаются разделители: точка, тире, знак вопроса и т.д., а также выражения в кавычках или апострофах, являющиеся названиями различных объектов.

Определение. Пусть *Cb* – размеченный концептуальный базис вида (6.2.4). Тогда текстообразующей системой (т.о.с.), согласованной с базисом *Cb*, называется произвольный упорядоченный набор *Tform* вида

$$(Morphbs, Constr, infconstr, Markers) \quad , \qquad (6.3.4)$$

где *Morphbs* - морфологический базис Р-типа вида (6.3.3), *Constr* – счетное множество символов, не пересекающееся с множеством словоформ *W* , *infconstr* – отображение из множества *Constr* в первичный информационный универсум *X(B),* где *B* – концептуальный базис, являющийся первым компонентом р.к.б. *Cb, Markers* – конечное множество символов, не пересекающееся с множествами *W* и *Constr,* и выполняются следующие условия: (а) для каждого *d* из множества *Constr* элемент *tp(infconstr(d))* является сортом из множества *St(B);* (б) множества *W, Constr, Markers* не включают апострофы и кавычки.

Элементы множеств *W, Constr* и *Markers* называются соответственно словоформами (или словами), конструктами и разделителями (или маркерами) системы *Tform.*

Очевидно, если задан *Morphbs* - морфологический базис Р-типа вида (6.3.3) , то заданы, в частности, алфавит А и множество словоформ W.

Определение. Пусть *Tform* – текстообразующая система вида (6.3.4). Тогда: *Names (Tform) = Names1 ∪ Names2*, где *Names1*={*‘x’*| *x* – цепочка в алфавите *A*}, *Names2*={*“y”*| *y* – цепочка в алфавите *A*}; *Textunits(Tform) = W ∪ Constr ∪ Names(Tform) ∪ Markers*; *Texts(Tform)* – множество всех конечных последовательностей вида $d_1,\ldots, d_n$, где $n \geq 1$, для $k=1,\ldots, n$ $d_k \in$ *Textunits(Tform).*

Определение. Пусть *Cb* – размеченный концептуальный базис, *Tform* – текстообразующая система вида (6.3.4), согласованная с базисом *Cb.* Тогда отображение *tclass* из *Textunits(Tform)* в *Parts(Morphbs) ∪{констр, имя}* и отображение *subclass* из *Textunits(Tform)* в *Subparts(Morphbs) ∪{nil}*, где *nil* –пустой элемент, задаются следующими соотношениями: (1) если *u∈ W(Tform),* то *tclass(u) = prt(u);* если *u ∈ Constr,* то *tclass(u) = констр;* если *u∈ Names(Tform),* то *tclass(u) = имя; ;* если *u∈ Markers,* то *tclass(u) =маркер*; (2) если *u∈ W(Tform),* то *subclass(u) = subprt(u);* если *u ∈ Constr,* то *subclass(u)= tp(infconstr(u)),* где *infconstr* и *tp* – отображения, являющиеся соответственно компонентами текстообразующей системы и первичного информационного универсума *X(B(Cb))*; если *u∈ Names(Tform) ∪ Markers,* то *subclass(u)= nil.*

6.4. Понятие лексико-семантического словаря

Рассмотрим модель словаря, ставящего в соответствие единицам текстов (“контейнеров”, “поступили” и далее др.) единицы семантического (или, другими словами, информационного) уровня; такие единицы в лингвистике называют *семами*. Лексико-семантический словарь (л.с.с.) является одним из основных компонентов ЛБД. Часть информационных единиц, соответствующих словоформам, мы будем считать символами; они являются элементами первичного информационного универсума *X(B(Cb)),* где *Cb* – размеченный концептуальный базис (р.к.б.), построенный для выбранной области, *B* – концептуальный базис (к.б.), являющийся первым компонентом *Cb*. Примеры таких единиц:

опубликование, *поступление1*, *станция1*, *станция2* и т.д. Другая часть информационных единиц имеет определенную структуру. Например, с прилагательным *"алюминиевый"* из *W* можно связать выражение *Материал (z, алюминий).*

Определение. Пусть *S* – сортовая система (с.с.) вида (2.5.1.). Тогда *семантической размерностью* системы S называется наибольшее такое число $k>1$, что найдутся сорта $u_1,\ldots,u_k \in St$, такие, что для любых $i,j=1,\ldots,k$ при $i \neq j$ u_i и u_j сравнимы для *отношения совместимости Tol* (т. е. $(u_i, u_j) \in Tol$). Это число *k* обозначается через *dim(S)* .

Таким образом, *dim(S)* - это наибольшее количество различных *"семантических осей"*, используемых для описания одной сущности в рассматриваемой области.

Пример. Рассмотрим понятия *"фирма"* и *"институт"*. Можно выделить три *семантических контекста* использования слов, соответствующих эти понятиям. Во-первых, фирма или институт могут разрабатывать прибор, технологию и т.д., поэтому в предложениях с этими словами может быть реализована *семантическая координата "интеллектуальная система"*. Во-вторых, мы можем сказать: "Эта фирма расположена возле м. Таганская," - тогда в этой фразе реализуется *семантическая координата "пространственный объект"*. Наконец, фирмы, институты имеют руководителя. Например, мы можем сказать: "Директор этой фирмы - А. Н. Семенов." В данной фразе реализована *семантическая координата "организация"*.

Мы будем предполагать в рассматриваемых примерах, что семантическая размерность используемых сортовых систем равна четырем или трем.

С содержательной точки зрения, под *лексико-семантическим словарем* мы будем понимать некоторое конечное множество *Lsdic*, состоящее из упорядоченных наборов вида

$$(i, lec, pt, sem, st_1,\ldots,st_k, comment) \quad , \qquad (6.4.1),$$

где $i \geq 1$ - порядковый номер набора (нужен для организации циклов), а остальные компоненты интерпретируются следующим образом. Компонент *lec* является элементом множества лексем *Lecs* рассматриваемого морфологического базиса; *pt* -

обозначение части речи лексемы *lec*; компонент *sem* является цепочкой, обозначающей одно из возможных значений лексемы *lec*.

Компонент *sem* для глаголов, причастий, деепричастий является информационной единицей, связанной с соответствующим отглагольным существительным. Например, глагол *"поступить"* имеет два значения: (1) поступление абитуриента в учебное заведение; (2) поступление физического объекта на какой-то пространственный объект (например, товара на склад). Поэтому, в частности, началом одного из наборов возможного лексико-семантическим словаря будет последовательность элементов i_1, *поступить*, *глаг*, *поступление1*, а началом другого набора - последовательность i_2, *поступить*, *глаг*, *поступление2*.

Число k является семантической размерностью рассматриваемой сортовой системы, т.е. $k = dim(S(B(Cb))$, где Cb – рассматриваемый размеченный концептуальный базис; $st_1,\ldots, st_k$ - различные *семантические координаты* сущности, характеризуемой понятием *sem*. Например, если *sem=фирма*, то st_1=*интс*, st_2=*простр.об*, st_3=*орг*, k= 3. Если же сущность, характеризуемая понятием *sem*, имеет различные семантические координаты $st_1,\ldots,st_p$, где $p<k$, то $st_{p+1},\ldots,st_k$ - это специальный *пустой элемент nil*. Компонент *comment* является пояснением на естественном языке смысла понятия *sem* либо пустым элементом *nil*.

Определение. Пусть *Cb* - размеченный концептуальный базис вида (6.2.4), *Morphbs* - морфологический базис вида (6.3.3), *Qmk* – разметка вопросов вида (6.2.1), первичный информационный универсум *X(B(Cb))* и множество переменных *V(B(Cb))* не включают символ *nil* (пустая сема). Тогда *лексико-семантическим словарем (л.с.с.)*, согласованным с р.к.б. *Cb* и морфологическим базисом *Morphbs*, называется произвольное конечное множество *Lsdic*, состоящее из упорядоченных наборов вида (6.4.1.), где $i \geq 1$, $lec \in Lecs$, $pt=prt(lec)$, $sem \in Lp(Cb) \cup \{nil\}$, $k=dim(S(B(Cb)))$; для каждого $p=1,\ldots, k$ $st_p \in St(B(Cb)) \cup \{nil\}$, $comment \in A^+ \cup \{nil\}$ и выполняются следующие условия:

(а) никакие два набора из *Lsdic* не могут иметь один и тот же первый компонент *i*; (бг) если два набора из *Lsdic* имеют разные значения компонента *sem*, то эти два набора имеют разные значения компонента *comment.*

Пример. *Lsdic* может быть определён так, что *Lsdic* включает следующие наборы:

(112, контейнер, сущ, контейнер1, дин. физ. объект, nil, nil, *“ёмкость”),*

(208, поступить, глаг, поступление1, ↑cит, nil, nil, *«поступить в вуз»),*

(209, поступить, глаг, поступление2, ↑cит, nil, nil, *«поступил груз»),*

*(311, алюминевый, прилаг, Материал(*z1, *алюминий),физ.об,* nil, nil, nil*),*

*(358, зеленый, прилаг, Цвет(*z1, *зелен), физ.об,* nil, nil, nil*),*

*(411, пассажирский, прилаг, Назначение(*z1, *перемещение1 * (Объект1, опред множ * (Кач-состав, человек))), дин.физ.об,* nil, nil, nil*),*

*(450, Италия, сущ, нек страна * (Назв, ‘Италия’), простр.об,* nil, nil, *«страна”),*

*(512, обувь, сущ, нек множ * (Кач-состав, * изделие1 * (Вид, обувн)), дин.физ.об,* nil, nil, *«термин, обозначающий различные множества обувных изделий»).*

6.5. Словари глагольно-предложных семантико-синтаксических фреймов

Ключевую роль в формировании предложений играют глаголы, причастия, деепричастия и отглагольные существительные, выражая разнообразные отношения между объектами рассматриваемой предметной области.

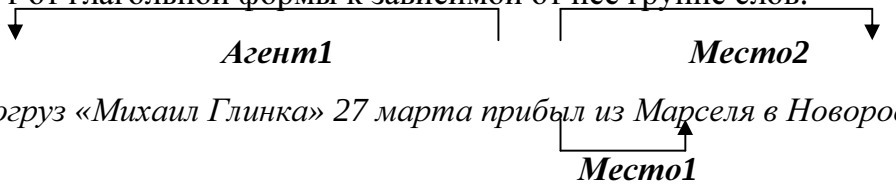
Тематической ролью называется смысловое отношение между значением глагольной формы (личной формы глагола, неопределенной формы глагола, причастия, деепричастия, отглагольного существительного) и значением зависящей от нее в предложении группы слов. Тематические роли называются также концептуальными падежами, семантическими падежами, глубинными падежами и семантическими ролями.

Впервые понятие глубинного падежа было предложено американским лингвистом Ч. Филлмором в 1968 году. Это понятие очень быстро стало широко популярным в компьютерной лингвистике, поскольку лежит в основе базовых

процедур установления смысловых отношений между значением глагольной формы и значением зависящей от нее группы слов.

Пример 1. Пусть Т1 = «Сухогруз «Михаил Глинка» 27 марта прибыл из Марселя в Новороссийск». Глагол прибыл обозначает некоторое событие вида прибытие, с которым можно связать метку e1 (event1). В Т1 упоминаются (называются) следующие объекты: некоторый корабль x1; некоторый город x2 с названием “Марсель”; некоторый город x3 с названием “Новороссийск”.

В событии *e1* объект *x1* играет роль «Агент-действия» (*Агент1*), *x2* играет роль «Исходное место для движения» (*Место1*), *x3* играет роль «Целевое место» (*Место2*). Тогда говорят, что в Т1 реализуются тематические роли *Агент1*, *Место1*, *Место2*, а также тематическая роль Время. Эта информация представляется размеченным текстом Т1, где стрелки с обозначением тематических ролей ведут от глагольной формы к зависимой от нее группе слов:



Пример 2. Пусть Т2 = «Сухогруз «Михаил Глинка» прибыл из Марселя.».

В Т2 явно реализуются только тематические роли *Агент1*, *Место1*, а тематические роли *Время* и *Место2* только подразумеваются в силу семантики глагола *прибывать*. Таким образом, фраза с одним и тем же глаголом в одном и том же значении могут явно выражать разные подмножества тематических ролей.

На формальном уровне мы будем интерпретировать тематические роли, как названия бинарных отношений, где первым атрибутом является ситуация, а вторым – реальный или абстрактный объект, играющий определенную роль в этой ситуации. При этом, если элемент $r \in R_2(B)$, где B – концептуальный базис, интерпретируется как тематическая роль, то его тип $tp(r)$= - это цепочка вида $\{(s,u)\}$, где s – конкретизация выделенного сорта *сит* (ситуация), а u – сорт из $St(B)$.

Словари глагольно-предложных фреймов содержат такие шаблоны (фреймы), которые позволяют представлять необходимые условия для реализации конкретной тематической роли в сочетании *Глагольная форма + Предлог + Зависимая группа*

слов, где *предлог* может быть пустым *(nil)*, а *зависимая группа слов* является либо существительным с зависимыми словами или без них, либо конструктом, то есть числовым значением параметра. Например, такими сочетаниями являются выражения “прибыть в порт”, “выехал из города”, “подготовить 4 статьи”, “купила итальянские туфли”, “ вернулся до 16:30”.

Определение. Пусть *Cb* – размеченный концептуальный базис вида (4.2.4), *Tform* – текстообразующая система вида (6.3.4), согласованная с *Cb*, *Morphbs* – морфологический базис вида (6.3.3) , *Lsdic* – лексико-семантический словарь, согласованный с *Cb* и *Tform*. Тогда словарем глагольно-предложных фреймов (с.г.п.ф.), согласованным с *Cb, Tform* и *Lsdic*, называется произвольное конечное множество *Vfr*, состоящее из упорядоченных наборов вида:

(k, semsit, form, refl, vc, sprep, grcase, str, trole, expl) (6.5.1)

где $k \geq 1$, $semsit \in X(B)$, $form \in \{неопр, личн, nil\}$, $refl \in \{возвр, невз, nil\}$, $vc \in \{действ, страд, nil\}$, $sprep \in W \cup \{nil\}$, где *nil* – пустой элемент, *W* - множество словоформ системы *Tform;* если $sprep \in W$, то $prt(sprep)=предлог$; $0 \leq grcase \leq 6$, $str \in St(B)$, *trole* – бинарный реляционный символ из первичного информационного универсума *X(B)*, причем $tp(trole)=\{(s,u)\}$, где $s, u \in St(B)$, причем *s* – конкретизация сорта “ситуация” *сит* (т.е. *сит* -> *s*); $expl \in A^{+} \cup \{nil\}$.

Компоненты произвольного набора вида (6.5.1) из *Vfr* интерпретируются следующим образом: *k* – порядковый номер набора; *semsit* – семантическая единица, обозначающая вид ситуации (прибытие, отлет, получение и др.); *form* - признак формы глагола, *неопр* – указатель неопределенной формы глагола; *личн* – указатель личной формы глагола, т.е. глагола в изъявительном или сослагательном наклонении; *refl* - признак возвратности глагола или причастия, *возвр* - указатель возвратной формы, *невз* - указатель невозвратной формы; *действ*, *страд* – признаки действительного и страдательного залога.

Компоненты *semsit, form, refl, vc* задают требования к глагольной форме, а компоненты *sprep, grcase, str* – требования к слову или группе слов, зависящей от глагольной формы и выражающей вместе с ней тематическую роль *trole*. Цепочка *sprep* – предлог, в том числе и составной (например, «в течение»), или nil; grcase –

код грамматического падежа от 1 до 6, либо 0 – как указатель отсутствия такой информации; *str* – семантическое ограничение на значение зависимой группы слов или слова; trole – та тематическая роль, необходимое условие реализации которой представляет данный набор (фрейм); *expl* – пример на ЕЯ, поясняющий тематическую роль, либо пустой пример nil.

Пример. Построим словарь. *Vfr1*, позволяющий устанавливать смысловые отношения в предложениях с глаголами *подготовить* и *поступить*. С каждым из этих глаголов будем связывать два значения. С глаголом *подготовить* – значения *подготовка1* (подготовка отчета, статьи и т.д.) и *подготовка2* (подготовка мастеров спорта и т.д.); с глаголом *поступить* – значения *поступление1* (поступление абитуриента в вуз) и *поступление2* (поступление контейнера на склад и т.д.).

Словарь *Vfr1* должен быть полезен, в частности, для семантического анализа текстов Т1 = “Профессор Семенов подготовил в июне отчет для НИИ «Заря»”; Т2 = “Профессором Семеновым в июне был подготовлен отчет для НИИ «Заря»”; Т3 = “Профессор Семенов подготовил в течение 1995-2003 годов трех кандидатов химических наук”; Т4 = “Контейнер поступил на склад в среду”;
Т5 = “В 1999 году Игорь поступил в МИЭМ”.

Можно построить такой р.к.б. *Cb*, что его первым компонентом является к.б. *B*, и выполняются следующие предположения: *St(B)* включает подмножество
{орг, интс, мом, инф.об, дин.физ.об, сит, квалиф, соб, простр.об, строка};
X(B) включает подмножество *{сейчас, иссл.инст, вуз, профессор, канд.хим.н, уч.завед, чел, ‘Семенов’, подготовка1, подготовка2, МИЭМ, контейнер1, склад1, отчет1, Назв, Фам, Месяц, июнь, среда, 3, 1995, 2002, 2003, Квал, Агент1, Объект1, Объект2, Продукт1, Время, Место1, Место2, Адресат1, Уч.зав}*;
сит → соб (т.е. событие является частным случаем ситуации);
tp (подготовка1) = tp (подготовка2) = tp (поступление1) =
tp (поступление2) = ↑соб ; *tp (иссл.инст) = tp (вуз) = ↑орг*простр.об*интс* ;
*tp (чел) = ↑интс*дин.физ.об* ; *tp (склад1) = ↑простр.об* ;
tp (контейнер1) = ↑дин.физ.об ; *tp (профессор) = tp (канд.хим.н) = квалиф* ;

*tp (МИЭМ) = орг*простр.об*интс tp (Агент1) = {(соб, интс)}* ;
tp (Адресат1) = {(соб, орг)}, tp (Время) = {(соб, мом)} ;
tp (Место1) = tp (Место2) = {(соб, простр.об)} ; *tp (Фам) = {(интс, строка)};*
tp ('Семенов') = строка ; *tp (Объект1) = {(соб, дин.физ.об)}* .

Тогда пусть *Vfr1* – множество, состоящее из следующих упорядоченных наборов:

k semsit form refl vc sprep grcase str trole expl

(1, подготовка1, личн, нвз, действ, nil, *1, интс, Агент1, 'И.П.Сомов подготовил (учебное пособие)')* ,
(2, подготовка1, личн, нвз, страд, nil,*5, интс, Агент1, 'Профессором Семеновым была подготовлена (книга)')* ,
(3, подготовка1, личн, нвз, действ, nil, *4, инф.об, Продукт1, '(И.П.Сомов) подготовил книгу'),*
(4, подготовка1, личн, нвз, страд, nil, *1, инф.об, 1, инф.об, 'Статья была подготовлена (за три недели)')* ,
(5, подготовка2, личн, нвз, действ, nil, *4, квалиф, Объект2, '(школа) подготовила 5 мастеров спорта')* ,
(6, подготовка1, nil, nil, *в, 0, мом, Время, 'подготовил в 2001-м году')* ,
(7, подготовка2, nil, nil, *в, 0, мом, Время, 'подготовит в 2003-м году')* ,
(8, поступление1, личн, действ, в, 4, орг, Уч.зав, '(Игорь) поступил в МИЭМ') ,
(9, поступление1, личн, действ, nil, *1, интс, Агент1, 'Игорь поступил (в МГУ)'),*
(10, поступление2, личн, нвз, действ, nil, *1, дин.физ.об, Объект1, 'контейнер поступил (на склад)')* ,
(11, поступление2, личн, нвз, действ, на, 4, простр.об, Место2, '(контейнер) поступил на склад') ,

(12, поступление2, личн, нвз, действ, в, 4, простр.об, Место2, ‘(контейнер) поступил вчера в магазин’) .

6.6. Формализация необходимых условий реализации данного смыслового отношения в сочетаниях вида “Глагол + Зависимая группа слов”

Пусть Tform – текстообразующая система, согласованная с размеченным концептуальным базисом (р.к.б.) Cb вида (6.2.4); $T \in$ Texts(Tform), длина (T) = nt, $1 \leq$ posn1 $\leq$ nt, posn1 – позиция существительного из T, $1 \leq$ posvb $\leq$ nt, posvb - позиция глагола из T; sem1, sem2 $\in$ X(B), где X(B) – первичный информационный универсум концептуального базиса B, являющегося первым компонентом р.к.б. Cb; prep - предлог из W или пустой предлог nil, $1 \leq$ grcase ≤ 6, rel – бинарный реляционный символ из X(B), интерпретируемый как название тематической роли. Тогда условимся, что запись

(T, posn1, sem1, prep, grcase, posvb, sem2, rel) $\in$ *Смысл-связь1*

интерпретируется следующим образом: если с элементом t_{posn1} , т.е. с существительным в позиции posn1, можно связать семантическую единицу sem1 и грамматический падеж с кодом grcase, к этому существительному в тексте T относится предлог prep (в частности, пустой предлог *nil*), с элементом t_{posvb} , т.е. с глаголом в позиции posvb, можно связать семантическую единицу sem2, то между элементами t_{posvb} и t_{posn1} может существовать смысловое отношение, являющееся тематической ролью с именем rel .

Пример 1. Предположим, что редакция некоторого научного журнала использует в своей работе интеллектуальную информационно-поисковую систему (ИПС), и этой системе задан вопрос B1 = “Когда поступила статья профессора Сомова?”. Если проставить после каждой текстообразующей единицы из B1 ее порядковый номер, то получится следующее (более удобное для анализа) представление вопроса B1: “Когда (1) поступила (2) статья (3) профессора (4) Сомова (5) ? 6)”.

Допустим, что лигвистическая база данных (ЛБД) редакционной ИПС включает компоненты, формальными моделями которых являются некоторый лексико-семантический словарь Lsdic и некоторый словарь глагольно-предложных фреймов Vfr, согласованные с размеченным концептуальным базисом (р.к.б.) Cb, причем первой составляющей Cb является концептуальный базис B. Пусть словарь Lsdic включает наборы

(k, поступить, глаг, поступление1, соб, nil, nil, 'поступление в вуз'),

(k+1, поступить, глаг, поступление2, соб, nil, nil, 'поступление контейнера на склад'),

(m, статья, сущ, статья1, инф.об, дин.физ.об, nil, 'статья, отправленная вчера в газету'),

(m+1, статья, сущ, статья2, инф.об, nil, nil, 'статья как часть юридического документа').

Тогда, очевидно, в вопросе B1 реализуется значение глагола "поступить", которому в словаре Lsdic соответствует семантическая единица *поступление2*, и реализуется значение существительного "статья", которому соответствует семантическая единица *статья1*.

Обозначим через *Объект1* тематическую роль, реализующуюся (в контексте вопроса B1) в сочетании "поступила статья", и предположим, что первичный информационный универсум X(B) включает бинарный реляционный символ *Объект1*. Заметим, что словоформа "статья" в вопросе B1 не связана с каким-либо предлогом, т.е. этой словоформе соответствует пустой предлог *nil*.

Пусть posn1 = 3 (позиция в вопросе B1 слова "статья") , posvb = 2 (позиция в вопросе B1 слова " поступила "). Слово "статья" в тексте B1 находится в именительном падеже, кодом которого является число 1. Тогда, с учетом сделанных предположений, выполняется соотношение

$$(B1, 3, статья1, nil, 1, 2, поступление2, Объект1) \in Смысл\text{-}связь1 .$$

Используя формальные средства, определим более точно смысл соотношения $(T, posn1, sem1, prep, grcase, posvb, sem2, rel) \in Смысл\text{-}связь1$.

Определение. Если Tform – текстообразующая система вида (6.3.4), Morphbs – морфологический базис Р-типа вида (6.3.3), то пусть

Nouns (Tform) = {d ∈ W │ prt(d) = сущ } , Prepositions (Tform) = {d ∈ W │ prt(d) = предлог } , Verbs (Tform) = {d ∈ W │ prt(d) = глаг}.

Таким образом, *Nouns(Tform), Prepositions(Tform)* и *Verbs(Tform)* – соответственно множества существительных, предлогов и глаголов, задаваемых текстообразующей системой Tform.

Определение. Пусть *Cb* - размеченный концептуальный базис вида (6.2.4), где первой составляющей *Cb* является концептуальный базис *B; Tform* – текстообразующая система, согласованная с р.к.б. *Cb*; *X(B)* – первичный информационный универсум базиса *B* ; m – семантическая размерность сортовой системы *S(B)* ; $R_2(B)$ – подмножество *X(B),* состоящее из всех бинарных редяционных символов; *Lsdic* – лексико-семантический словарь (л.с.с.), согласованный с р.к.б. *Cb*, *Vfr* - словарь глагольно-предложных фреймов, согласованный с р.к.б. *Cb* и л.с.с. *Lsdic*; *N* – множество положительных целых чисел.

Тогда подмножество *Смысл-связь1* декартова произведения

Texts(Tform) x N x X(B) x Prepositions (Tform) x {1,…, 6} x N x X(B) x $R_2(B)$

задается следующим условием:

упорядоченный набор *(T, posn1, sem1, prep, grcase, posvb, sem2, rel)* принадлежит множеству *Смысл-связь1* ⇔ когда *T ∈ Texts(Tform),* длина *(T) = nt, 1 ≤ posn1 ≤ nt,* t_{posn1} *∈ Nouns (Tform), 1 ≤ posvb ≤ nt,* t_{posvb} *∈ Verbs (Tform), prep ∈ Prepositions (Tform) U { nil }, 1 ≤ grcase ≤ 6, sem1, sem2 ∈ X(B), rel ∈* $R_2(B)$, т.е. *rel* – бинарный реляционный символ из первичного информационного универсума *X(B),*

существуют такие наборы (фреймы) *Fr1, Fr2* из *Lsdic* соответственно видов *(i1, lec1, сущ , sem1,* s_1*,…,* s_m*, comment1) ,*

(i2, lec2, глаг , sem2, st_1*,…,* st_m*, comment2)* и существует такой фрейм *Fr3* из *Vfr* вида *(k1, semsit, form, refl, vc, relat, sprep, grc, str, expl) ,*

что выполняется каждое из следующих условий:

Условие 1*: lcs (* t_{posn1}*) = lec1 , lcs (*t_{posvb} *) = lec2, semsit = sem2,*

sprep = prep , grc = grcase, relat = rel.

Условие 2: *grcase ∈ Падежи(morph(t_{posn1}) ,* где – множество числовых кодов всех грамматических падежей, которые могут соответствовать существительному в позиции *posn1*.

Условие 3: Пусть для произвольного сорта *s ∈ St Gener(s) = {u ∈ St(B) │ (u, s) ∈ Gen(B)}* , т.е. *Gener(s)* - множество всех сортов, являющихся обобщениями сорта s , включая сам сорт *s*. Тогда *str* входит в объединение множеств *Gener(s_i)* по всем сортам s_1,…, s_m , являющихся компонентами фрейма *Fr1* из *Lsdic* и отличных от пустого сорта *nil* .

Условие 4: Глагол имеет значение признака возвратности *refl,* значение формы *form* и значение залога *vc* .

Легко видеть, что смысл условия 3 заключается в том, что найдется такое *p, 1 ≤ p ≤ m,* что сорт s_p является конкретизацией сорта *str* – компонента глагольно-предложного фрейма *Fr3*.

Пример 2. Вернемся к размеченному представлению вопроса В1 “Когда (1) поступила (2) статья (3) профессора (4) Сомова (5) ? 6)”.

Допустим, что словарь глагольно-предложных фреймов Vfr включает набор вида *(n, поступление2, личн, нвз, действ, дин.физ.об, nil, 4, Объект1, ‘поступил контейнер’)* , где n – порядковый номер набора, личн – признак личной формы глагола (в отличие от неопределенной формы), действ - признак действительного залога глагола, дин.физ.об – сорт “динамический физический объект”, 1 – числовой код именительного падежа.

Пусть справедливы следующие соотношения (см. Пример 1): *posn1 = 3, posvb = 2, sem1 = статья1, sem2 = Lsdic[k+1].sem = поступление2, str = Lsdic[m].st_2 = дин.физ.об, prep = nil , = grcasy = 1, = rel = Объект1, lcs (статья) = статья , lcs (поступила) = поступить*. Тогда имеет место соотношение

(B1, posn1, sem1, prep, grcase, posvb, sem2, rel) ∈ Смысл-связь1 ,

т.е. *(B1, 3, статья1, nil, 1, 2, поступление2, Объект1) ∈ Смысл-связь1* ,

где *3 = posn1* – позиция слова “статья”, *2 = posvb* - позиция слова “поступила”.

Смысловые связи в предложении могут существовать не только между глаголами и существительными, но и между глаголами и конструктами, т.е. числовыми значениями различных параметров. Как и в случае существительных, на вид смысловой связи между глаголом и конструктом влият предлог, который может находиться перед конструктом. Например, в сочетаниях “нагрейти воду до 12 градусов” и “нагрейти воду на 12 градусов” реализуются разные смысловые отношения.

Пусть Tform – текстообразующая система, согласованная с размеченным концептуальным базисом.*Cb* вида (4.2.4); $T \in$ Texts(Tform), длина (T) = nt, $1 \leq$ posc1 $\leq$ nt, posc1 – позиция конструкта из T (т.е. $t_{posc1} \in$ Constr(Tform)), $1 \leq$ posvb $\leq$ nt, posvb - позиция глагола из T; semvb - семантическая единица из первичного информационного универсума X(B), соответствущая отглагольному существительному, образованному от глагола в позиции posvb; prep - предлог из W или пустой предлог nil, rel – бинарный реляционный символ из X(B), интерпретируемый как название тематической роли. Тогда будем считать, что запись

(T, posnc1, prep, posvb, semvb, rel)) $\in$ *Смысл-связь2*

интерпретируется следующим образом: если к конструкту в позиции posc1 относится предлог prep, то между элементами t_{posvb} и t_{posc1} может существовать смысловое отношение, являющееся тематической ролью с именем rel .

Пример 3. Рассмотрим предписания T1 = “ Нагрейти воду до 18 градусов” и T2 = “Нагрейти воду на 18 градусов”. Заменим эти предписания их размеченными представлениями “Нагрейти (1) воду (2) до (3) 18 градусов (4)” и “Нагрейти (1) воду (2) на (3) 18 градусов (4)”.

Пусть posc1 = 4, posvb = 1 , semvb = нагревание . Тогда можно определить размеченный концептуальный базис Cb, словари Lsdic и Vfr, согласованные с р.к.б. Cb, и отношение *Смысл-связь2* так, что будут выполняться соотношения

(T1, posnc1, до , posvb, semvb, Предельное-значение) $\in$ *Смысл-связь2* ,

(T2, posnc1, на , posvb, semvb, Приращение-значения) $\in$ *Смысл-связь2* .

6.7. Словари предложных семантическо-синтаксических фреймов

6.7.1. Формальное определение словаря предложных фреймов

Рассмотрим следующую проблему: каким образом в сочетании “Существительное1 + Предлог + Существительное2” или “Существительное1 + Существительное2” установить, какое именно смысловое отношение реализуется в этом сочетании. Рассмотрим идею решения на примерах.

Пример 1. Пусть С1 = ”Поезд из Праги”. Со словом “поезд” связано понятие *поезд1*, а этому понятию соответствует сорт *дин.физ.об* (“динамический физический объект”).. Словоформа “Праги” обозначает город. С понятием *город1* связан сорт *простр.об* (пространственный объект). Существительное “Прага” находится в родительном падеже, тогда можно представить, что лингвистическая база данных (ЛБД) включает семантическо-синтаксический шаблон вида

(k1, ‘из’, дин.физ.об, простр. об., 2, Место3, ‘посылка из Таганрога’) ,

смысл которого заключается в следующем: *k1* – номер шаблона; ‘из’ – предлог; *дин.физ.об*. – сорт “динамический физический объект”, связанный с первым существительным; *простр.об* – сорт, связанный со вторым существительным; 2 – код родительного падежа, причём второе существительное должно находиться в родительном падеже; *Место3* – обозначение смыслового отношения, которое реализуется в сочетании “Существительное1 + ‘из’ + Существительное2” при выполнении заданных условий; *‘посылка из Таганрога’* – пример выражения, в котором реализуется отношение *Место3*. Вместо *Место3* мы могли бы написать *Исходный-пространственный-объект*.

Легко видеть, что сочетание С1 совместимо с этим шаблоном, имеющим номер *k1*.

Пример 2. Пусть С2=”статья в журнале”, тогда ЛБД может включать шаблон вида

(k2, ‘в’, инф.об, инф.об, 6, Место4, ‘глава в книге’) ,

где *k2* – номер шаблона; ‘в’ – предлог; *инф. об*. – сорт “информационный объект”, 6 – код предложного падежа.

Со словом “статья” связаны понятия *статья1*, *статья2*. Понятие *статья1* интерпретируется как понятие, которому соответствует выражение “статья в журнале, газете и т.д.”; *статья2* – это отдельная смысловая часть документа (юридическое понятие). Сорт “информационный объект” связан с каждым из этих понятий. Поэтому выражение С2 совместимо с шаблоном, имеющим номер *k2*. Следовательно, в выражении С2 может реализовываться смысловое отношение *Место4*.

Пример 3. Пусть С3=”статья профессора”, и ЛБД включает шаблон вида

(k3, nil, инф.об, интс, 2, Авторы, ‘поэма Пушкина’) ,

где *nil* – пустой предлог; *интс* – сорт “интеллектуальная система”, 2 – код родительного падежа. С лексемой “профессор” можно связать сорт *интс* и сорт *дин.физ.об* (“динамический физический объект”). Поэтому сочетание С3 совместимо с шаблоном, имеющим номер *k3*.

Определение. Пусть *Cb* – размеченный концептуальный базис вида (6.2.4), *B=B(Cb), Morphbs* – морфологический базис вида (6.3.3); *Tform* - текстообразующая система вида (6.3.4), согласованная с р.к.б. *Cb*; *Lsdic* – лексико-семантический словарь, состоящий из записей вида (6.4.1), согласованный с *Cb* и *Tform*. Тогда словарём предложных семантико-синтаксических фреймов, согласованным с *Cb, Tform и Lsdic* , называется произвольное конечное множество *Frp*, состоящее из упорядоченных наборов вида

(i, prep, sr1, sr2, grc, rel, ex) (6.7.1)

где $i \geq 1$; $prep \in Lecs \cup \{nil\}$, *где nil* – цепочка, обозначающая пустой предлог; если $prep \in Lecs$, то *prt (prep) = предлог*; $sr1, sr2 \in St(B)$; $1 \leq grc \leq 6$; $rel \in R_2(B)$; $R_2(B)$ – множество бинарных реляционных символов, являющееся подмножеством первичного информационного универсума *X(B(Cb))*; $ex \in A+$.

Компоненты набора вида (6.7.1) из множества *Frp* интерпретируются следующим образом. Натуральное число $i \geq 1$ является порядковым номером набора (используется для организации циклов), *prep* – это предлог из множества лексем *Lecs* или пустой предлог *nil*. Элементы *sr1* и *sr2* интерпретируются как сорта, которые можно связать соответственно с

первым существительным и вторым существительным в лингвистически правильном сочетании “Сущ.1 + prep + Сущ.2”; *grc (grammatic case)* – код падежа, в котором должно находиться второе существительное в таком правильном сочетании; *rel* – обозначение смыслового отношения, которое может реализовываться в таком сочетании при выполнении указанных условий; *ex* – пример выражения, в котором реализуется то же самое отношение rel.

Пример. Можно построить такие размеченный концептуальный базис *Cb*, морфологический базис *Morphbs*, текстообразующую систему *Tform*, лексико-семантический словарь *Lsdic*, и словарь предложных семантико-синтаксических фреймов *Frp*, что *Frp* включает семантические шаблоны (фреймы): с номерами *k1, k2, k3,* рассмотренные в примерах 1 - 3, а также следующие шаблоны:

(k4, ‘от’, вещество, болезнь, 2, Против1, ‘таблетки от гриппа’);

(k5, ‘от’, вещество, дин.физ.об, 2, Против2, ‘мазь от комаров’);

(k6, ‘от’, физическое явление, физ.об, 2, Эффект1, ‘тень от дома’).

Потребуем, чтобы выполнялось следующее условие: в словаре *Frp* не найдётся таких наборов, в которых совпадают компоненты *prep ≠ nil, sr1, sr2, grc*, но не совпадают компоненты *rel* или *ex*. В таком случае четвёрка *(prep, sr1, sr2, grc)* однозначно определяет смысловое отношение *rel*.

6.7.2. Формализация необходимых условий существования определенного смыслового отношения в сочетании из двух существительных с учетом предлога

Определение. Пусть *B* – произвольный концептуальный базис (к.б.). Тогда для произвольного сорта $s \in St(B)$ $Gener(s)=\{u \in St(B) \mid (u,s) \in Gen(B)\}$, т.е. *Gener(s)* – множество всех сортов, являющихся обобщением сорта *s*, включая сам сорт s.

Определение. Пусть *Tform* – текстообразующая система (т.о.с.), согласованная с размеченным концептуальным базисом *Cb* вида (6.2.4.), *Morphbs* – морфологический базис Р-типа вида (6.3.3), $Nouns(Tform)=\{ d \in W \mid prt(d)=сущ\}$,

Prepositions(Tform)={ d∈W | prt(d)=предлог}, X(B) – первичный информационный универсум концептуального базиса *B=B(Cb), m* – семантическая размерность сортовой системы *S(B), $R_2(B)$* – подмножество *X(B),* состоящее из всех бинарных реляционных символов, N^+ - множество положительных целых чисел, *T* – текст из *Texts(Tform).* Тогда подмножество *Смысл-связь3* множества

Texts(Tform) ×N ×X(B) ×Prepositions(Tform) ×{1, … 6} ×N^+ ×X(B) ×R_2(B)

задается следующим условием:

(T, posn1, sem1, prep, grcase, posn2, sem2, rel)∈Смысл-связь3 ⇔

T∈Texts(Tform), 1 ≤posn1 < posn2 ≤длина (T), sem1, sem2 ∈X(B),

prep∈Prepostions(Tform) U {nil}, где *nil* – пустой предлог, *1 ≤ grcase ≤ 6, rel ∈ R_2(B),* и существуют такие фреймы *Fr1, Fr2 из Lsdic* соответственно видов

(i1, lec1, сущ, sem1, s_1, …, s_m, comment1) ,

(i2, lec2, сущ, sem2, st_1, …, st_m, comment2),

а также фрейм *Fr* из *Frp* вида *(k1, prep, sr1, sr2, grc, rel, ex),*

что *lcs(t_{posn1})=lec1, lcs(t_{posn2})=lec2, grc ∈ Падежи (fmorph(t_{posn2})),*

sr1 входит в объединение множеств *Gener(s_i)* по всем сортам s_1, …, s_m, отличным от *nil, sr2* входит в объединение множеств *Gener(st_i)* по всем сортам st_1,…, st_m, отличным от *nil*.

Пример. Проиллюстрируем применение определения отношения *Смысл-связь3* к проверке возможности реализации смыслового отношения *Против1* в сочетании “микстура от кашля”, являющегося фрагментом предложения Т1= “В аптеке # 18 продается новая микстура от кашля”.

Пусть множество сортов *St(B)* включает элементы *вещество, жидк.вещество, (вещество, жидк.вещество) ∈ Gen(B), Lsdic* включает элементы *Fr1, Fr2* соотвественно видов

(n1, микстура, микстура1, жидк.вещество, nil, nil, ‘лекарство’)

(n2, кашель, кашель1, болезнь, nil, nil, ‘вид заболевания’).

Пусть *posn1=6, posn2=8.* Тогда с учетом того, что словарь *Frp* включает набор (*k4, ’от’, вещество, болезнь, 2, Против1, ‘таблетки от гриппа’),* справедливо соотношение

(T, posn1, микстура1, от, 2, posn2, кашель1, Против1) ∈ Смысл-связь3.

6.8. Лингвистические базисы

6.8.1. Формализация семантической информации, связанной с вопросительными словами

Определим понятие системы вопросительных словосочетаний. Будем называть ролевыми вопросительными словосочетаниями пары вида *(qw, d),* где *qw* – это предлог, либо пустой предлог nil; $d \in W$ - некоторое слово, являющееся либо вопросительно-относительным местоимением, либо местоименным наречием. Например, такими сочетаниями являются пары *(nil, кто)* , *(nil, кому)* , *(для, кого)* , *(у, кого)* , *(nil, откуда)* .

Наша языковая интуиция позволяет связать с каждой из таких пар некоторое достаточно общее смысловое отношение:

(nil, кто) → Агент; (nil, кому) → Адресат; (для, кого) → Адресат;

(у, кого) → Источник1 (частные случаи: Продавец, Поставщик).

В связи с этим в состав лингвистической базы данных введём ещё один словарь.

Определение. Пусть выполняются предположения из определения словаря предложных семантико-синтаксических фреймов. Тогда системой ролевых вопросительных словосочетаний, согласованной с размеченным концептуальным базисом *Cb*, морфологическим базисом *Morphbs* и лексико-семантико-синтаксическим словарём *Lsdic* , называется произвольное конечное множество, состоящее из упорядоченных наборов вида

$$(i, prep, qw, relq) \quad , \qquad (6.8.1)$$

где $i \geq 1$, $prep \in Lecs \cup \{nil\}$, $qw \in W$, $prt(qw) \in \{местоим, наречие\}$, $relq \in R_2(B(Cb))$; в случае $prep \neq nil$ $prt(prep) = предлог$, $subprt(qw) = вопр\text{-}относ\text{-}местоим$; в случае $prep = nil$ $subprt(qw) \in \{вопр\text{-}относ\text{-}местоим$, $местоим\text{-}наречие\}$.

Пример. Определим *B, Cb, Morphbs, Lsdic, Rqs* так, что *Rqs* включает наборы

(1, nil, кто, Агент) , (2, nil, кому, Адресат) ,

(3, для, кого, Адресат), (4, у, кого, Источник1) , (5, на, чём, Инструмент)

(6, nil, когда, Время) , (7, nil, откуда, Место1) , (8, nil, куда, Место2).

(3, для, кого, Адресат), (4, у, кого, Источник1) , (5, на, чём, Инструмент).

6.8.2. Понятие лингвистического базиса

Лингвистические базисы являются формальными моделями лингвистических баз данных (ЛБД)..

Определение. Упорядоченный набор *Lingb* вида

$$(Cb, Tform, Lsdic, Vfr, Frp, Rqs) \qquad (6.8.2)$$

называется *лингвистическим базисом (л.б.)* ↔ когда *Cb* – размеченный концептуальный базис (р.к.б.) вида (6.2.4), *Tform* – текстообразующая система (т.о.с.) вида (6.3.4), согласованная с р.к.б. *Cb, Lsdic* – лексико-семантический словарь (л.с.с.), согласованный с р.к.б. *Cb* и т.о.с. *Tform*, *Vfr* – словарь глагольно-предложных семантико-синтаксических фреймов, согласованный с р.к.б *Cb*, т.о.с. *Tform*, л.с.с.*Lsdic*; *Rqs* – система вопросительных словосочетаний, согласованная с *Cb*, *Tform*, *Lsdic*.

Формальное понятие лингвистического базиса отражает наиболее существенные черты логической структуры широко применимых ЛБД. Это понятие конструктивно в том смысле, что на его основе можно проектировать ЛБД практически полезных лингвистических процессоров.

Понятие лингвистического базиса обобщает научные результаты автора, опубликованные в работах (Фомичев 1978б, 1979, 1980, 1986а, 1987б, 1988ж, 1990г, 1991б; Fomichov 1992, 2002a; Fomichov, Kochanov 2001).

Глава 7
НОВЫЙ МЕТОД ВЫПОЛНЕНИЯ ПРЕОБРАЗОВАНИЯ “ЕЯ-ТЕКСТ ➔СЕМАНТИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ”

В этой главе предложена новая структура данных (названная матричным семантико-синтаксическим представлением текста), используемая в качестве промежуточной формы отображения семантической структуры входного ЕЯ-текста, для последующего построения СП текста. Описывается новый предметно-независимый метод преобразования входного ЕЯ-текста (вопроса, команды, сообщения) из подъязыка русского языка в его СП, являющееся выражением некоторого СК-языка.

Структуры данных, ассоциированные с текстом в рамках заданного лингвистического базиса

В следующем параграфе будет описан новый метод преобразования ЕЯ текста в его СП; метод использует новую форму представления промежуточных результатов анализа текста. - матричное семантико-синтаксическое представление (МССП) текста. Для определения МССП ЕЯ-текста опишем ряд вспомогательных структур данных, которые ассоциированы с входным текстом прикладной интеллектуальной системы в рамках рассматриваемого лингвистического базиса.

7.1.1. Компонентно-морфологическое представление текста

Морфологическое представление. Временно пропуская ряд математических деталей, условимся под морфологическим представлением текста *T* длины *nt* понимать двумерный массив *Rm* с индексами столбцов *base* и *morph*, элементы которого интерпретируются следующим образом. Пусть *nmr* – количество строк массива *Rm*, построенного для текста *T*, и *k* – номер строки массива *Rm*, т.е. $1 \le k \le nmr$. Тогда *Rm[k, base]* – это лексема, соответствующая некоторой словоформе в позиции *p* из текста *T*. При тех же предположениях *Rm[k, morph]* – это последовательность наборов значений морфологических признаков, соответствующих словоформе в позиции *p*.

Определение. Пусть *Tform* – текстообразующая система вида (6.3.4), *Morphbs* – морфологический базис вида (6.3.3), $T \in Texts(Tform)$, *nt* – длина (*T*). Тогда *морфологическим представлением текста T* называется двумерный массив *Rm* с индексами столбцов base и morph, для которого выполняются следующие условия:

1. Каждая строка массива *Rm* содержит информацию о какой-то словоформе из входного текста *T*, т.е., если *nmr* – количество строк массива *Rm*, то для каждого *i* от *1* до *nmr* в тексте *T* найдется такая позиция *p*, $1 \le p \le nt$, что $t_p \in W$, $Rm[i, base] = lcs(t_p)$, $Rm[i, morph] = fmorph(t_p)$.
2. Для каждой словоформы из текста *T* найдется строка в *Rm*, представляющая морфологическую информацию об этой словоформе, т.е. для каждой позиции *p* в тексте *T*, где $1 \le p \le nt$, $t_p \in W$, найдется такое *k*, где $1 \le k \le nmr$, что $lcs(t_p) = Rm[k, base]$, $fmorph(t_p) = Rm[k, morph]$.
3. Любые две строки массива *Rm* различаются либо значением лексемы в столбце *base*, либо набором значений морфологических признаков в столбце morph, т.е. если $1 \le k \le nmr$, $1 \le q \le nmr$, $k \ne q$, то либо $Rm[k, base] \ne Rm[q, base]$, либо $Rm[k, morph] \ne Rm[q, morph]$.

Таким образом, произвольная строка массива Rm указывает лексему (базовую форму) и совокупность наборов значений морфологических признаков, связанных

с какой-то лексической единицей из текста Т. В то же время для каждой лексической единицы из Т найдена соответствующая строка в Rm.

Пример. Пусть Т1= «В(1) каком(2) московском(3) издательстве(4) в(5) 2001-м(6) году(6) вышла(7) работа(8) по(9) искусственному(10) интеллекту(10) «Основы обработки знаний» (11) профессора(12) Сомова(13)?(14)» (после каждого элементарного выражения из текста указан порядковый номер элементарной значащей единицы текста, к которой относится данное выражение). Тогда морфологическое представление Rm текста Т1 будет иметь следующую форму:

base	morph
в	md_1
какой	md_2
московский	md_3
издательство	md_4
выходить	md_5
работа	md_6
по	md_7
искусственный интеллект	md_8
профессор	md_9
Сомов	md_{10}

Рис. 7.1. Структура морфологического представления Rm

Здесь md_1,…, md_{10} – числовые коды наборов морфологических признаков, связанных с соответствующими словами из входного текста Т1. В частности, md_4 кодирует следующие сведения: часть речи – существительное, подкласс части речи - существительное нарицательное, число – единственное, падеж – предложный. Набор неотрицательных целых чисел md_{10} кодирует следующую информацию: часть речи - существительное, подкласс часть речи – существительное собственное, падеж - родительный, число - единственное.

Классифицирующее представление. Пусть *Tform* – текстообразующая система вида (4.3.4), *T∈Texts(Tform), nt* – длина (*T*). Тогда, с содержательной точки зрения, классифицирующим представлением текста T, согласованным с морфологическим представлением Rm текста T, будет называться двумерный массив Rc с количеством строк nt и индексами столбцов unit, tclass, subclass, mcoord, элементы которого интерпретируются следующим образом.

Пусть k – номер произвольной строки массива Rc, т.е. 1≤k≤nt. Тогда Rc[k, unit] является одной из элементарных значащих единиц текста T , т.е если $T=t_1 \ldots t_{nt}$, то найдется такая позиция p, 1≤p≤nt, что Rc[k, unit]= t_p. Если Rc[k, unit] – словоформа, то Rc[k, tclass], Rc[k, subclass], Rc[k, mcoord] являются соответственно обозначениями части речи, подкласса части речи, последовательности наборов значений морфологических признаков.

Если Rc[k, unit] – конструкт (т.е. числовое значение параметра), то Rc[k, tclass] – цепочка *констр*, Rc[k, subclass] – сорт информационной единицы, соответствующей данному конструкту, Rc[k, mcoord]=0.

Пример. Пусть T1= «В каком московском издательстве в 2001-м году вышла работа по искусственному интеллекту «Основы обработки знаний» профессора Сомова?» Тогда классифицирующее представление *Rc* для текста T1,согласованное с *Rm*, может иметь следующий вид:

unit	tclass	subclass	mcoord
в	предлог	nil	1
каком	местоим	вопрос-относит-местоим	2
московском	прилаг	nil	3
издательстве	сущ	сущ-нарицат	4
в	предлог	nil	1
2001-м году	констр	момент	0
вышла	глаг	глаг-в-изъявит-накл	5
работа	сущ		6

по	предлог	сущ-нарицат	7
искусственному	сущ	nil	8
интеллекту		сущ-нарицат	
«Основы	имя		
обработки			
знаний»	сущ		9
профессора	сущ	сущ-нарицат	10
Сомова	маркер	сущ-собств	0
?			

Рис. 7.2. Структура классифицирующего представления Rc

Определение. Пусть Tform – т.о.с. вида (6.3.4), Morphbs – м.б. вида (6.3.3), nt – длина (T), T∈ Texts(Tform), Rm - морфологическое представление T. Тогда классифицирующим представлением текста T, согласованным с Rm, назовём двумерный массив Rc с индексами столбцов unit, tclass, subclass, mcoord и количеством строк nt, для которого выполняются следующие условия:

1. Для k = 1,.., nt Rc[k,unit] = tk.
2. Если 1≤k≤nt, tk ∈ W, то Rc[k, tclass] = prt(tk),

Rc[k, subclass] = subprt(tk), и найдется такое q, 1≤q≤nrm, где nrm количество строк в Rm, что Rc[k, mcoord]=q, fmorph(tk)=les(tk)=Rm[q,base], Rm[q,morph]

3. Если 1≤k≤nt, tk∈ Constr, Rc[k, tclass]=констр, Rc[k, subclass]=tp((tk)), Rc[k,mcoord]=0.
4. Если 1 ≤ k ≤ t, t_k∈ Names (Tform), то Rc[k, tclass]=имя, Rc[k, subclass]=ml, Rc[k, mcoord]=0.
5. Если 1≤k≤nt, tk∈ Markers, то Rc[k, tclass]=Маркер, Rc[k, subclass]=ml, Rc[k,mcoord]=0.

Таким образом, классифицирующее представление текста T задаёт следующие сведения:

1. Для каждой лексической единицы указывает часть речи, подкласс части речи (если он определён) и номер строки из морфологического представления Rm, перечисляющей числовые коды морфологических признаков, соответствующих данной лексической единице..

2. Для каждого конструкта задает класс *констр* и подкласс, являющийся сортом информационной единицы, соответствующей конструкту.

3. Для каждого элемента из множества Names(Tform) указывает класс *имя*, подкласс nil и число 0 в столбце mcoord.

4. Для каждого разделителя (знаки препинания) указывается класс *маркер*, подкласс nil и 0 в столбце mcoord.

Определение. Пусть *Tform* – т.о.с. вида (6.3.4), *T∈ Texts(Tform).* Тогда компонентно-морфологическим представлением текста *T* будем называть упорядоченную пару вида *(Rm, Rc),* где *Rm* – морфологическое представление текста *T*, *Rc* – классифицирующее представление текста *T,* согласованное с *Rm.*

7.1.2. Проекции компонентов лингвистического базиса на входной текст

Пусть *Lingb* – л.б. вида (6.8.2), и *Dic* – какой–либо из следующих компонентов *Lingb:* лексико-семантический словарь (л.с.с.) *Lsdic,* словарь глагольно-предложных фреймов *Vfr,* словарь предложных фреймов *Frp.* Тогда проекцией *Dic* на входной текст *T∈ Texts(Tform)* назовем двумерный массив, строки которого представляют всю информацию из *Dic,* которая относится к лексическим единицам из *T.* Вводимые ниже определения позволяют уточнить эту идею.

Определение. Пусть *Lingb* – л.б. вида (6.8.2), *T∈ Texts (Tform), nt* – длина *T, (Rm, Rc)* – компонентно-морфологическое представление *T.* Тогда проекцией л.с.с. *Lsdic* на входной текст *T* назовем двумерный массив *Arls* с индексами столбцов *ord, sem,* st_1,…, st_m, *comment,* где m - семантическая размерность сортовой системы S(B(Cb(Lingb))), удовлетворяющий следующим условиям:

1. Элементами столбца ord являются порядковые номера элементарных значащих единиц текста T, т.е. номера строк классифицирующего представления Rc.
2. Элементы столбцов *sem*, st_1,…, st_m, *comment* интерпретируются так же, как и одноименные компоненты л.с.с. *Lsdic* , т.е. *sem* – простая или составная семантическая единица, st_1,…, st_m – несравнимые для отношения совместимости *Tol* элементы сортового множества *St(B(Cb(Lingb)))*, *comment* – естественно-языковое описание смысла единицы *sem*.
3. Для каждой словоформы wd из W, входящей в строку $q \geq 1$ и столбец *unit* классифицирующего представления *Rc*, найдутся такая строка с номером $k \geq 1$ в массиве *Arls* и такой набор вида *(i, lec, pt, sem,* st_1,…, st_k, *comment)* из лексико-семантического словаря *Lsdic* , где $i \geq 1$, что компоненты этого набора *sem*, st_1,…, st_m, *comment* совпадают с элементами одноименных столбцов из строки *k*; *Rc[q, tclass] = pt*, и, если *Rc[q, mcoord] = m*, то *Rm[m, base] = lec*.
4. Строки массива *Arls*, в столбце *sem* которых расположены семантические единицы, представляющие разные значения одной и той же лексемы, следуют подряд и не могут перемежаться строками, в столбце *sem* которых отражены возможные значения каких-то других лексем.
5. Пусть для произвольной строки с номером *n* массива *Arls* элемент *q = Arls[n, ord]* является номером какой-то строки классифицирующего представления *Rc*, т.е. $1 \leq q \leq nt$. Тогда элементы столбцов *sem*, st_1,…, st_m, *comment* строки с номером *n* массива *Arls* совпадают с одноименными компонентами какого-то набора вида (*j, lec, pt, sem,* st_1,…, st_m, *comment*) из л.с.с. *Lsdic* , где $lcs(t_q) = lec$ - лексема, $prt(t_q) = pt$ – обозначение части речи.
6. В массиве *Arls* нет повторяющихся строк.

Легко видеть, что из сформулированных выше условий вытекает, что количество строк narls массива Arls равно сумме количеств значений лексем, соответствующих словоформам из входного текста T.

Пример. Пусть T1= «В(1) каком(2) московском(3) издательстве(4) в(5) 2001-м(6) году(6) вышла(7) работа(8) по(9) искусственному(10) интеллекту(10) «Основы(11) обработки(11) знаний(11)» профессора(12) Сомова(13)?(14)». Тогда массив Arls для T1 может иметь следующий вид:

N	**ord**	**sem**	**st1**	**st2**	**st3**	**st4**	**comment**
1	3	Город(z, Москва)	Простр .об	nil	nil	nil	nil
2	7	Издательство	орг	интс	простр.об	nil	nil
3	7	Выход1	сит	nil	nil	nil	‘Игорь вышел из комнаты’
4	7	Выход2	сит	nil	nil	nil	‘Книга вышла в 1988 году’
5	8	Работа1	соб	nil	nil	nil	‘Эта работа заняла 4 часа’
6	7	Работа2	инф.об	дин.физ. об	nil	nil	‘Работа профессора Новикова была отправлена экспресс-почтой’
7	10	Иск. интеллект	Науч. обл	nil	nil	nil	‘Научное направление «искусств. интеллект»’
8	12	Нек.чел *(Квалиф, ‘профессор’)	интс	дин.физ. об.	nil	nil	nil
9	13	Нек.чел * (Фам., ‘Сомов’)	интс	дин.физ. об	nil	nil	nil

Рис. 7.3. Структура массива Arls

Смысл рассмотрения двумерного массива Arvfr, называемого проекцией словаря глагольно-предложных фреймов Vfr на текст T, заключается в следующем: для каждой глагольной формы из текста T в этом массиве размещаются все шаблоны (фреймы) из словаря *Vfr,* позволяющие находить возможные смысловые отношения между значением данной глагольной формы и значением зависящей от нее в предложении из текста T группы слов.

Определение. Пусть Lingb – л.б. вида (6.8.2), T∈Texts(Tform), nt = длина(T), (Rm, Rc) – компонентно-морфологическое представление T, Arls – проекция лексико-семантического словаря Lsdic на текст T. Тогда назовем двухмерный массив Arvfr с индексами столбцов *nb, semsit, form, refl, vc, sprep, grc, str, trole, example* проекцией словаря глагольно-предложных фреймов Vfr на текст T, если выполняются следующие условия:

1. Элементами столбца *nb* являются порядковые номера элементарных значащих единиц текста T, являющихся глагольными формами (глаголами или причастиями), т.е. номера строк классифицирующего представления Rc.
2. Элементы столбцов *semsit, form, refl, vc, sprep, grc, str, trole, example* интерпретируются так же, как и одноименные компоненты словаря глагольно-предложных фреймов Vfr .
3. Пусть $q \geq 1$ – номер произвольной строки классифицирующего представления *Rc*, для которой Rc[q, tclass] ∈ {*глаг, прич*} (т.е. *q* – позиция произвольной глагольной формы из входного текста T); k - номер такой произвольной строки массиве *Arls*, что *Arls[k, ord] =q,* и *semunit* – элемент *Arls[k, sem]*. Тогда для каждого набора *d* из словаря *Vfr* вида

   (i, semsit, refl, form, vc, sprep, grc, str, trole, example) ,

   где $i \geq 1$ и *semsit = semunit* , в массиве *Arvfr* найдется такая строка с номером *m*, что *Arvfr [m, semsit] = semunit,* и элементы строки m, расположенные в столбцах *refl, form, vc, sprep, grc, str, trole, example,* совпадают с одноименными компонентами набора *d.*

Другими словами, для каждой глагольной формы t_q из текста T и любого значения *semunit* формы t_q массив *Arvfr* должен включать каждый глагольно-предложный фрейм из словаря *Vfr*, связанный со значением *semunit*.

4. Пусть *m* – номер произвольной строки массива *Arvfr* . Тогда найдутся такая строка массива Rc с номером q и такая строка массива Arls с номером k, что выполняются соотношения Rc[q, tclass] ∈ {*глаг*, *прич*}, *Arls[k, ord] =q*, *Arls[k, sem].= Arvfr [m, semsit]* (т.е. каждая строка массива *Arvfr* содержит глагольно-предложный фрейм из словаря Vfr, связанный с некоторым значением какой-либо глагольной формы из текста T).
5. Строки массива *Arvfr*, в столбце *semsit* которых расположены семантические единицы, представляющие разные значения одной и той же лексемы, следуют подряд и не могут перемежаться строками, в столбце *semsit* которых отражены возможные значения каких-то других лексем.
6. В массиве *Arvfr* нет повторяющихся строк

Пример. Вопрос T1 = “В каком московском издательстве в 2001-м году вышла работа по искусственному интеллекту “Основы обработки знаний” профессора Сомова?” включает глагольную форму “вышла”. Глагол “выходить” имеет, в частности, два значения, которым мы поставим в соответствие семантические единицы *выход1* и *выход 2*. Эти значения реализуются соответственно в предложениях “Теплоход вышел из порта в 8:30” и “Учебник вышел в 2003-м году”.

Будем считать, что со значением *выход1*связаны тематические роли Агент1 (Агент действия), Место1 (обозначение отношения между событием, заключающимся в перемещении физического объекта в пространстве, и исходным пространственным объектом), Место2 (обозначение отношения между событием, заключающимся в перемещении физического объекта в пространстве, и целевым пространственным объектом), Время, Длительность, Целевой-предмет (данное отношение реализуется, например, в предложении “Игорь вышел из дома за хлебом”).

Со значением *выход2* будем ассоциировать тематические роли Инф-объект (Информационный объект), Время, Организация (обозначение отношения между событием, заключающимся в опубликовании информационного объекта, и организацией, опубликовавшей этот объект).

Тогда массив *Arvfr*, построенный по тексту T1 с учетом массива Arls, рассмотренного в предыдущем примере, может иметь следующий вид:

nb	*semsit*	*fm*	*refl*	*vc*	*trole*	*sprep*	*grc*	*str*	*example*
6	Выход1	из	нвз	действ	Агент1	*nil*	1	дин. физ.об.	Он вышел (из дома)
6	выход1	из	нвз	действ	Место1	из	2	простр. об	(он) вышел из дома
6	выход1	из	нвз	действ	Длит	на	0	зн.длит.	Вышел на 2 часа
6	выход1	из	нвз	действ	Время	в	0	момент	(Теплоход) вышел (из порта) в 8:30
6	выход1	из	нвз	действ	Целе-вой пред-мет	за	5	Дин. Физ.об	Вышел (из дома) за хлебом
6	выход2	из	нвз	действ	Инф-объект	*nil*	1	инф.об.	(В издательстве “Белый город») вышел альбом
6	выход2	из	нвз	действ	Время	в	0	момент.	(Книга) вышла в 2002-м году
6	выход2	из	нвз	действ	Органи-зация	в	6	орг	(книга) вышла в издательстве

Рис. 7.4. Пример массива *Arvfr*

Связь с массивом *Arls* существляется через поле *semsit.* Шаблон из *Arvfr* связан со строкой с номером *k* массива *Arls*, если они относятся к одной и той же лексической единице из текста, и значение поля sem для массива *Arls* совпадает со значением поля *semsit* из массива *Arvfr*.

Аналогично строится массив *Arpfr* – проекция словаря предложных фреймов *Frp* на входной текст. Этот массив предназначен для отображения всех сведений из словаря Frp, относящихся к предлогам из текста T и к пустому предлогу *nil*.

Определение. Пусть Lingb – л.б. вида (6.8.2), T∈ Texts(Tform), nt = длина(T), (Rm, Rc) – компонентно-морфологическое представление T, Arls – проекция лексико-семантического словаря Lsdic на текст T. Тогда назовем двухмерный массив Arfrp с индексами столбцов *prep, sr1, sr2, grc, rel, ex* проекцией словаря предложных фреймов Frp на текст T, если выполняются следующие условия:

1. Пусть $q \geq 1$ – номер произвольной строки классифицирующего представления *Rc*, для которой Rc[q, tclass] = предлог (т.е. *q* – позиция произвольного предлога из входного текста T), и *pr = Rc[q, unit]*. Тогда в массиве Arfrp нйдется такая строка с номером $k \geq 1$, что Arfrp[k, prep] = pr, и в словаре Frp найдется набор вида (6.7.1), в котором prep] = pr, и компоненты *sr1, sr2, grc, rel, ex* совпадают с элементами одноименных столбцов массива Arfrp, расположенными в строке *k*.
2. Пусть *k*. – номер произвольной строки массива *Arfrp.* Тогда найдутся такие строка массива Rc с номером $q\ (1 \leq q \leq nt)$ и набор *d* вида (6.7.1) в словаре Frp, что *Arfrp[k, prep] = Rc[q, unit], Rc[q, tclass] = предлог*, и компоненты *sr1, sr2, grc, rel, ex* набора *d* совпадают с элементами одноименных столбцов массива Arfrp, расположенными в строке *k*.
3. Пусть *d* – произвольный набор из словаря предложных фреймов Frp, для которого компонентом *prep* является пустой предлог *nil*, и *h* - набор, получающийся из *d* удалением первого компонента (порядкового номера) набора. Тогда компоненты *sr1, sr2, grc, rel, ex* набора *h* совпадают с элементами одноименных столбцов некоторой строки массива Arfrp (т.е. Arfrp включает все шаблоны, или фреймы, позволяющие находить

возможные смысловые отношения в сочетаниях вида “Существительное1 + пустой предлог + Существительное2”).

4. Все строки массива Arfrp различны.

Пример. В тексте T1 = “В каком московском издательстве в 2001-м году вышла работа по искусственному интеллекту “Основы обработки знаний” профессора Сомова?” встречается предлог “по”. Этот предлог может использоваться, в частности, в сочетаниях “прогулка по городу” и “книга по физике”, причем во втором случае реализуется то же смысловое отношение, что и в тексте T2. Поэтому массив Arfrp для T1 может иметь вид, представленный на рисунке 7.5.

prep	sr1	sr2	grc	rel	ex
по	соб	простр.об	3	Место3	Прогулка по парку
по	инф.об	обл.деят	6	Область1	Книга по живописи

Рис. 7.5. Пример фрагмента массива Arfrp

7.2. Матричное семантико-синтаксическое представление ЕЯ – текста

Рассмотрим новую структуру данных, предлагаемую в данной работе в качестве промежуточной формы представления результатов семантико-синтаксического анализа ЕЯ-текстов и называемую матричным семантико-синтаксическим представлением (МССП) входного текста T.

МССП ЕЯ-текста T – это строково-числовая матрица Matr с индексами столбцов *locunit, nval, prep, posdir, reldir, mark, qt, nattr, contr*, позволяющая по информации о возможных видах коротких сочетаний слов найти смысловые отношения между элементами предложения T, а также указать одно из нескольких возможных значений каждой лексической единицы.

Таким образом, МССП текста – это матрица Matr следующего вида:

locunit	*nval*	*prep*	*posdir*	*reldir*	*mark*	*qt*	*nattr*	*contr*

Рис. 7.6. Структура матричного семантико-синтаксического представления текста

Количество строк матрицы Matr равно *nt* - количеству количество строк в классифицирующем представлении *Rc* , т.е. количеству выделенных элементарных значащих единиц текста.

В столбце ***locunit*** (*location of unit,* место единицы) указывается наименьший номер строки массива *Arls*, которая соответствует лексической единице с порядковым номером *k,* где *k* – это номер строки массива *Rc* и номер строки матрицы *Matr*. Массив *Arls* представляет все наборы из лексико-семантического словаря (л.с.с.) *Lsdic*, которые содержат информацию о лексических единицах из входного текста. Массив *Arls* выше был назван проекцией л.с.с. *Lsdic* на входной текст *T*.

Можно сказать, что значение поля *locunit* для *k*-той единицы текста является координатой входа по этой единице в массив *Arls*.

Столбец ***nval*** *(number of values,* количество значений*)* в начальный момент построения *Matr* указывает количество всех строк из *Arls,* соответствующих *k*-й лексической единице, где *k* – номер строки *Rc* и *Matr*. После завершения построения *Matr* в столбце *nval* на пересечении с каждой строкой, соответствующей лексической единице, должно находиться значение 1, поскольку для каждой лексической единицы было найдено одно из нескольких возможных значений.

Столбец ***prep*** *(preposition,* предлог*)* для каждой строки с номер *k* указывает предлог (возможно, пустой предлог *nil*), относящийся к *k*-й лексической единице.

Рассмотрим назначение группы столбцов ***posdir*** *(posdir*$_1$*, posdir*$_2$*, …, posdir*$_n$), где *n* – константа в пределах от 1 до 10, зависящая от программной реализации. Пусть $1 \le d \le n$. Тогда будем использовать обозначение *Matr[k, posdir, d]* для элемента, расположенного на пересечении строки k и столбца из группы ***posdir*** с порядковым номером *d* в данной группе.

Если $1 \le k \le nt$, $1 \le d \le n$, то *Matr[k, posdir, d] = m*, где *m* – это либо 0, либо порядковый номер *d*-й лексической единицы из входного текста *T*, управляющей единицей с порядковым номером *k*. Для глаголов в главном предложении в этих столбцах стоит 0, т.к. для них нет управляющей единицы. Условимся считать, что существительное управляет стоящими перед ним прилагательными, а также относящимся к нему числом или количественным числительным (например, в сочетании “5 научных статей”). В группе столбцов *reldir* содержатся обозначения смысловых отношений, отраженных в группе столбцов *posdir*.

Рассмотрим соотношения, являющиеся исходными для заполнения столбцов *posdir* и *reldir* **.** Эти соотношения базируются на определениях отношений Смысл.связь1, Смысл.связь2, Смысл.связь3 из подраздела 6.7.2. Начнем с соотношений для сочетаний вида “Глагольная форма + Предлог + Существительное”, где Предлог может быть пустым предлогом *nil*.

Пусть $1 \le k \le nt$, $1 \le d \le n$, *Rc[k, tclass] = сущ*, $1 \le posvb \le nt$, $k \ne posvb$, *Rc[posvb, tclass]*$\in$*{глаг, прич}, Matr[k, posdir, d]=posvb, Matr[k,locunit]=loc1, Matr[k, nval]=1, Matr[posvb, locunit]=loc2, Matr[posvb, nval]=1, Arls[loc1, sem]=sem1, Arls[loc2, sem]=sem2, Matr[k, reldir, d]=relcat, prep1=Matr[k,prep]* .

Тогда найдется такой код грамматического падежа *grcase*, где $1 \le grcase \le 6$, что *(T, k,sem1, prep1, grcase, posvb,sem2, relat)*$\in$ *Смысл.связь1*. При этом *grcase*$\in$ *Падежи(Rm[j, morph]), где j=Rc[k,mcoord], Падежи(Rm[j, morph])* – множество числовых кодов всех грамматических падежей, указанных в каком-либо наборе морфологических признаков из *Rm[j, morph]*.

Если $1 \le k \le nt$, *Rc[k,tclass] = констр* (т. е. *k*-я единица текста является конструктом), и *Matr[k,reldir, d]=rel*, где *rel* – некоторый бинарный реляционный символ из *X(B)*, то выполняются следующие соотношения:

(1) найдется такое целое posvb, где $1 \le posvb \le nt$, что $Matr[k,posdir, d]=posvb$

(2) $Rc[posvb, tclass] \in \{глаг, прич\}$; (3) Если $Matr[k,prep]=prep1$, $Matr[posvb, locunit]=locvb$, $Arls[locvb,sem]=sem1$,

то $(T, k, prep1, posvb, sem1, rel) \in Смысл.связь2$.

Таким образом, будем считать, что управляющей единицей текста для каждого конструкта (т.е. представления значения числового параметра) является либо глагол, либо причастие).

Если $1 \le posn1 < posn2 \le n2$, $Rc\ [posn1, fclass] = Rc\ [posn2, fclass] = сущ$, $Matr[posn2, posdir] = posn1$, $Matr[posn2, prep] = prep1$, $Matr[posn1, reldir] = rel$, то элемент *rel* является именем смыслового соотношения между существительными в позициях posn1 и posn2, где слово в позиции posn1 управляет словом в позиции posn2 с учетом предлога prep1, относящегося ко второму существительному.

В этом случае должно выполняться следующее условие: если

$Matr[posn1, locunit] = loc1$, $Matr[posn1, nval] = 1$,

$Arls[loc1, sem] = sem1$, $Matr[posn2, locunit] = loc2$,

$Matr[posn2, nval] = 1$, $Arls[loc2, sem] = sem2$,

то найдется такое целое число grcase, где $1 \le grcase \le 6$, что $(T, posn1, sem1, prep,. Grcase, Posn2, sem2, rel) \in Смысл.связь3$, причем $grcase \in Cases(fmorph(Rc[posn2, unit))$.

Столбец ***mark*** (метка) предназначен для хранения переменных, обозначающих различные сущности из входного текста (в том числе события, на которые указывают глаголы, причастия, деепричастия, отглагольные существительные).

В столбце ***qt*** *(quantity)* – количество, помещается либо 0, либо число, которое указывается в тексте перед существительным и относится к существительным.

В столбце ***nattr*** *(number of attributes)* – количество атрибутов, указывается либо 0, либо количество прилагательных относящихся к существительному представленному в данной строке *k*, т.е. мы предполагаем, что $R_1[k].unit$ –это существительное.

В столбце ***contr*** (*control*, управление) помещается либо 0, либо число, позволяющее установить связь между главным предложением и причастным оборотом или придаточным предложением.

Пример. Пусть В1 = «Сколько контейнеров, поступивших в пятницу из Новороссийска, были отправлены АО “Радуга”?».

Тогда $k = 2$ – порядковый номер слова «контейнеров»; $p = 4$ – порядковый номер слова «поступивших», и *Matr[k, contr] = p, Matr[p, contr] =k.* Таким образом, если *k,* - позиция существительного, к которому ”прикреплено” причастие, то *Matr[k, contr]* – позиция этого причастия. Наоборот, если *p* - позиция причастия, то *Matr[p, contr]* – позиция существительного, к которому “прикреплено” это причастие.

Пусть П1 = ”Профессор Сомов работает в институте, который он закончил в 1978 году”, k = 5 (позиция словоформы “институте”), m = 9 (позиция словоформы “закончил”). Тогда *Matr[k, contr] = m,* и *Matr[m, contr] =k.*

Если придаточное определительное предложение соединено с главным предложением с помощью вопросительно-относительного местоимения в позиции j , то Matr[j, contr] = m, где m – позиция существительного из главного предложения, к которому прикреплено придаточное предложение.

Возможность использовать столбец contr в двух противоположных смыслах обусловлена тем, что каждая строка, соответствующая лексической единице, одназначно определяет ее часть речи.

Проиллюстрируем форму матричное семантико-синтаксическое представления (МССП) Matr.

Пример. Построим МССП текста Т1= «В(1) каком (2) московском (3) издательстве (4) в (5) 2001-м году (6) вышла (7) работа (8) по(9) искусственному интеллекту (10) ”Основы обработки знаний” (11) профессора (12) Сомова (13) ? (14)». В параграфе 4.9 для текста Т1 были построены массивы Arls, Arvfr, Arfrp. С учетом этого МССП Matr для текста Т1 может иметь следующий вид:

	Loc-unit	nval	prep	Pos-dir	reldir	Ma-rk	qt	nattr	Co-ntr
1	0	0	в	0, 0	nil, nil	nil	0	0	0
2	0	0	в	0, 0	Nil, nil	nil	0	0	0
3	1	1	в	4, 0	Место(z, Москва), nil	nil	0	0	0
4	2	1	в	7, 0	Простр.объект, nil	X1	0	1	0
5	0	0	в	0, 0	nil, nil	nil	0	0	0
6	0	0	в	7, 0	Время, nil	nil	0	0	0
7	3	1	nil	0, 0	nil, nil	L1	0	0	0
8	6	1	nil	7, 0	Объект 3, nil	X2	0	0	0
9	0	0	no	0, 0	nil, nil	nil	0	0	0
10	7	1	no	8, 0	Область 1, nil	X3	0	0	0
11	0	0	0	8, 0	Название, nil	nil	0	0	0
12	8	1	nil	8, 0	Авторы, nil	X4	0	0	0
13	9	1	nil	12, 0	Фамилия(z, ’Сомов’), nil	X4	0	0	0
14	0	0	0	0, 0	nil, nil	nil	0	0	0

Рис. 7.7. Пример матричного семантико-синтаксического представления текста

Построенная матрица отражает финальную конфигурацию МССП Matr. Это значит, что найдены все смысловые соотношения между единицами текста.

7.3. Новый метод преобразования ЕЯ-текстов в их семантические представления

7.3.1. Принципы установления соответствия между матричным семантико-синтаксическим представлением текста и его К-представлением

Как уже отмечалось выше, матричное семантико-синтаксическое представление (МССП) ЕЯ-текста Т строится как промежуточная структура для представления результатов семантико-семантического анализа Т. Следующий шаг должен заключаться в построении по МССП *Matr* некоторого К-представления текста Т, т.е. выражения некоторого стандартного К-языка, интерпретируемого как семантическое представление (СП) текста Т. В связи с этим ниже излагаются наиболее общие принципы преобразования МССП ЕЯ-текста Т в некоторое К-представление текста Т. На основе этих принципов в главе 9 разработан алгоритм преобразования МССП текста в его К-представление.

Рассмотрим структуры данных, позволяющие осуществить преобразование МССП текста в его К-представление.

Массив Sitdescr предназначен для построения семантических описаний ситуаций (в частности, событий), упоминаемых во входном тексте. Количество заполненных строк этого массива равно количеству глаголов и причастий в тексте. Столбец с индексом mrk хранит метки ситуаций (связь с Matr осуществляется через метки из этого столбца). Столбец с индексом expr предназначен для хранения семантических описаний ситуаций.

Рассмотрим пример заполнения массива Sitdescr. Пусть В1 = “На каких предприятиях, для которых поставляет картон АО “Старт”, выпускают мебель для кухни?”. Тогда массив Sitdescr может иметь следующий вид:

mrk	expr
e1	*Ситуация(e1, выпуск1 *(Агент1, нек множ *(Кач-состав, предприятие) : x1) (Объект1, нек множ * (Кач-состав, дин. физ. об. * (Класс1, мебель)(Цел место, нек кухня))))*
e2	*Ситуация(e2, поставка1 *(Агент1, нек орг *(Тип, АО)(Назв, "Старт") : x3) (Объект1, нек множ *(Кач-состав, дин.физ.об. *(Вещество, картон)) : x2)(Адресат, x1))*

Рис. 7.8. Пример конфигурации массива описания ситуаций Sitdescr

По такому массиву Sitdescr можно построить следующее КП *Semrepr* вида *Вопрос (x1, (Sitdescr[1] ∧ Sitdescr [2]))* , т.е. выражение

*Вопрос (x1, (Ситуация(e1, выпуск1 *(Агент1, нек множ *(Кач-состав, предприятие) : x1) (Объект1, нек множ * (Кач-состав, дин. физ. об. * (Класс1, мебель)(Цел место, нек кухня))))*

*∧ Ситуация(e2, поставка1 *(Агент1, нек орг *(Тип, АО)(Назв, "Старт") : x3) (Объект1, нек множ *(Кач-состав, дин.физ.об. *(Вещество, картон)) : x2)(Адресат, x1)))*

Начальным шагом формирования строки массива Sitdescr с меткой ситуации *ek* является построение выражения вида *Ситуация (ek, concept * ,* где *concept* – семантическая единица, квалифицирующая ситуацию и являющаяся значением поля *semnoun* массива Arls (проекции лексико-семантического словаря Lsdic на входной текст T) для строки, номер которой указан в столбце locunit матрицы Matr.

Например, в случае рассмотрения вопроса B1 первая и вторая строки массива Sitdescr получат соответственно значения *Ситуация(e1, выпуск1 * и Ситуация(e2, поставка1* **.**

После того, как сформировано начальное значение рассматриваемой строки массива Sitdescr, необходимо добавить в эту строку описания участников ситуации и соответствующие тематические роли. Для этого используется массив Performers

(“Исполнители-ролей”). Количество строк в этом массиве совпадает с количеством строк в классифицирующем представлении Rc и в МССП Matr

Наиболее простую структуру имеют семантические представления таких описаний участников ситуаций, которые являются сочетаниями вида “Существительное + Имя”, где Имя – это выражение в кавычках или апострофах. Пусть k – порядковый номер в тексте существительного из такого выражения, т.е. порядковый номер строки классифицирующего представления, соответствующей этому существительному. Тогда Performers [k] = *нек conc * (Назв, Имя)* ,
где *conc* – простое обозначение понятия, соответствующего данному существительному.

Например, для вопроса В1 Performers [9] = *нек акц-общ * (Назв, “Старт”)* ,
поскольку выражение АО занимает 9-е по порядку место в вопросе В1.

В лаконичной форме принципы заполнения массива Performers иллюстрирует слкдуюшая таблица. Первый столбец таблицы соответствует существительному с порядковым номером k, где $k \geq 1$; второй столбец – контексту для данного существительного, т.е. виду сочетания, в которое входит данное существительное; в третьем столбце указывается значение строки k массива Performers.

t_i	контекст	Performers [I]
контейнер	Поступил контейнер	*нек контейнер1*
контейнеры	Поступили контейнеры	*нек множ *(Кач-состав, контейнер1)*
контейнера	3 контейнера	*нек множ * (Колич, 3)(Кач-состав, контейнер1)*
контейнера	3 алюминиевых контейнера	*нек множ * (Колич, 3)(Кач-состав, контейнер1 * (Материал, алюм))*
контейнера	3 зел. алюмини-евых контейнера	*нек множ * (Колич, 3)(Кач-состав, контейнер1 * (Цвет, зел)(Материал, алюм))*

Ростов	Ростов	*нек город * (Назв, "Ростов")*
АО "Заря"	АО "Заря"	*нек орг *(Тип, АО)(Назв, "Заря")*
году	в 1998 году	*1998/nil/nil*
февраль	с февраля 1998	*1998/февраль/nil*
керамикой	с керамикой	нек множ *(Кач-состав, дин.физ.об. * (Вид, керамич-изделие))
обувью	с обувью	нек множ * (Кач-состав, дин.физ.об. * (Вид, обувн-изделие))
обувь	с обувью из Италии	нек множ *(Кач-состав, дин.физ.об. * (Вид, обувн изделие)(Изготовитель, нек страна * (Назв, "Италия")))
пластмассу	огнеустойчивую пластмассу	нек множ *(Кач-состав, дин. физ.об. * (Вещество, пластм *(Характ, огнеустойчив)).

Таблица 7.1. Виды элементов массива Performers

Пример. На основе указанных принципов по построенному в параграфе 7.2 МССП Matr текста T1 = "В каком московском издательстве в 2001-м году вышла работа по искусственному интеллекту «Основы обработки знаний» профессора Сомова ?" можно построить следующее К-представление (КП):

Вопрос (x1, Ситуация(e1, выход 1 (Простр.объект, нек изд-во1* (Место, Москва) : x21)(Время, 2001/год)(Объект3, нек работа2 * (Назв, 'Основы обработки знаний')(Область1, иск-интеллект)(Авторы, нек чел *(Квалификация, профессор)(Фамилия, 'Сомов') : x4) : x3))))) .*

Для фомирования массивов Performers и Sitdescr используются вспомогательные массивы Sembase и Semdes с количеством строк nt (количество единиц текста в класифицирующем представлении Rc входного текста T. Первоначально элементы этих масствов заполняются цепочкой nil.

Массив Sembase предназначен для явного отображения информации о семантике прилагательных и существительных из входного текста T.

Пример. Пусть В1 = “Какие европейские фармацевтические компании участвовали в выставке “ЭКСПОХИМ-2003”?”. Тогда в в класифицирующем представлении Rc вопроса В1 словоформам “европейские”, “фармацевтические”, “компании”, “выставке” будут соответствовать строки с номерами 2, 3, 4, 7. Лингвистический базис может быть определен таким образом, что в результате применения описываемого ниже алгоритма “Начало-постр-СемП” будут выполненв следующие операторы присваивания:

Sembase[2] := (Регион, Европа) ,

Sembase[3] := (Регион, Европа) (Область1, фармацевтика) ,

Sembase[4] := компания1 * (Регион, Европа) (Область1, фармацевтика) ,

Sembase[7] := выставка1 .

Цепочку Sembase[4] будем интерпретировать как составное обозначенияе понятия “европейская фармацевтическая компания”, но не как обозначение какой-то конкретной компании, цепочку Sembase[7] - как обозначение понятия “выставка”.

Массив Semdes предназначен для построения главных частей семантических представлений выражений из входного текста, обозначающих объекты или множества объектов (т.е. выражений, включающих существительные).

Пример. В контексте вопроса В1 выражение “европейские фармацевтические компании” обозначает некоторое конкретное множество компаний, а слово “выставка” – некоторую конкретную выставку. Поэтому

*Semdes[4] := нек множ * (Кач-состав, компания1 * (Регион, Европа) (Область1, фармацевтика)) ,*

*Semdes[7] := нек выставка1 * (Назв, 'ЭКСПОХИМ-2003’)* .

Здесь *нек* - информационная единица, интерпретируемая как квантор референтности (см. параграф 2.8). Поскольку выражения Semdes[4] и Semdes[7] начинаются с этой информационной единицы, постольку эти

выражения интерпретируются как обозначения конкретных сущностей, упоминаемых в тексте, а не как обозначения понятий.

Так как в позициях 2, 3 вопроса В1 расположены прилагательные, то Semdes[2] = Semdes[3] = nil .

В массиве Performers отличными от цепочки nil являются только элеиенты, соответствуюшие конструктам (числовым значениям параметров) или существительным. Если элемент в позиции k является конструктом, то Sembase[k] = Semdes[k] = nil , Performers [k] := Rc [k, unit] , где Rc - класифицирующее представление входного текста Т. Например, если позиции k соответствует значение цены учебника 112 рублей , то Performers [k] := 112/рубль .

Если же k – номер строки из Rc , соответствующей существительному, то Performers [k] := Semdes[k] + ' : v ‘ , где v – переменная, являющаяся меткой сущности в СП входного текста; здесь символ ‘+ ‘ будем интерпртировать как знак операции конкатенашии, т.е. операции приписывания справа к одной цепочке другой цепочки.

Пример. Для вопроса В1 *Performers [4] := Semdes[4] + ' : S1 ‘ ,*
Performers [7] := Semdes[7] + ' : x1 ‘ , т.е.
*Performers [4] := нек множ * (Кач-состав, компания1 * (Регион, Европа) (Область1, фармацевтика)) : S1 ,*
*Performers [7] := нек выставка1 * (Назв, 'ЭКСПОХИМ-2003’) : x1 .*

7.3.2. Формулировка метода

Введенные выше понятия и изложенные принципы позволяют сформулировать новый метод преобразования ЕЯ-текста (в частности, запроса, сообщения или команды) в СП текста. Эта метод предназначен для проектирования диалоговых систем и включает следующие три этапа преобразования:

Преобразование1: Компонентно-морфологический анализ входного текста..

Сущность преобразования заключается в следующем. По тексту Т на естественном языке строится одно или несколько компонентно-морфологических представлений (КМП) текста Т, т.е. один или несколько наборов вида (Rc, Rm), где Rc - классифицирующее представление текста и Rm – морфологическое представление текста, т.е. представление возможных значений морфологических признаков для тех компонентов текста Т, которые являются лексическими единицами (в отличие от числовых значений признаков, разделителей, выражений в кавычках или апострофах).

В большинстве случаев отдельным фразам из входного текста будет соответствовать единственное КМП. Если же либо входной текст Т неоднозначно разбивается на элементарные значащие единицы текста, либо неоднозначно определяется часть речи какой-либо единицы текста, то задаются уточнящие вопросы пользователю диалоговой системы, и неоднозначности снимаются после обработки ответов пользователя на эти вопросы.

Преобразование 2: Построение матричного семантико-синтаксического представления (МССП) текста.

Цель второго преобразования заключается в том, чтобы связать с каждым словом какое-то одно из возможных нескольких значений и в том, чтобы установить смысловые отношения между различными единицами текста.

Так как это делается постепенно, шаг за шагом, то МССП сначала является недоопределенным. Чтобы снять неоднозначности, могут задаваться уточняющие вопросы пользователю. Но, главным образом, используются сведения из лингвистической базы данных (ЛБД) о допустимых способах комбинирования разных единиц текста в лингвистически правильные сочетания.

Преобразование 3: Сборка семантического представления текста, являющегося К-представлением, по его МССП *Matr*.

Алгоритм, преобразующий МССП *Matr* в некоторое формальное выражение $Semrepr \in Ls(B)$, где *B* – концептуальный базис, являющийся первым компонентом используемого размеченного концептуального базиса (р.к.б.) *Cb*, *Ls(B)* – СК-язык в базисе *B*, будем называть *алгоритмом семантической сборки*.

7.3.3. Принципы выбора формы семантического представления для текстов различных видов

Форма семантического представления (СП) ЕЯ-текста Т, строящегося по МССП текста Т, должна зависеть от вида входного текста. Рассмотрим на примерах рекомендации по выбору формы СП, являющегося выражением стандартного К-языка (СК-языка) в некотором концептуальном базисе, т.е. К-представлением (КП) входного текста. В этих примерах СП входного текста Т будет являться значением строковой переменной Semrepr (Semantic representation).

Пример. Пусть Т1 = “Профессор Игорь Новиков преподает в Томске”. Тогда Semrepr = *Ситуация(e1, преподавание * (Время, #сейчас#)(Агент1, нек чел * (Квалиф, профессор)(Имя, ‘Игорь‘(Фамилия, 'Новиков’) : x2))(Место1, нек город * (Название, 'Томск’) : x3)).*

Пример. Пусть Т2 = “"Доставь ящик с деталями на склад № 3.". Тогда

Semrepr = *(Команда(#Оператор#, #Исполнитель#, #Сейчас#, e1) ∧ Цель (e1, доставка1*(Объект1, нек ящик * (Содерж1, нек множ * (Кач-состав, деталь)) : x1)(Место2, нек склад * (Номер, 3) : x2)) .*

Пример. Пусть Т3 = “Проходила ли в Азии международная научная конференция “COLING”?”. Тогда

Semrepr = *Вопрос(x1, (x1≡ Ист-знач (Ситуация (e1, прохождение2* (Время, нек мом * (Раньше ,#сейчас#) : t1)(Событие, нет конф* (Вид1, междун) (Вид2, научная) (Название, ‘COLING’) : x2) (Место, нек континент* (Название, ‘Азия’) :x3))))).*

Пример. Пусть Т4 = “Какое издательство опубликовало роман «Ветры Африки»?”. Тогда Semrepr = *Вопрос(x1, Ситуация(e1, опубликование * (Время, нек мом * (Раньше, #сейчас#) : t1) (Агент2, нек издательство: x1) (Объект3, нек роман1* (Название, ‘Ветры Африки’) :x3))) .*

Пример. Пусть Т5 = “ С какими зарубежными издательствами сотрудничает писатель Игорь Сомов?”. Тогда Semrepr = *Вопрос (S1, (Кач-состав (S1, издательство * (Вид-географич, зарубежное)) ∧ Описание (произв издательство**

*(Элем, S1) : y1, Ситуация(e1, сотрудничество * (Время, #сейчас#)(Агент1, нек чел* (Профессия, писатель)(Имя, 'Игорь')(Фамилия, 'Сомов'): x1)(Организация1, y1)))))* .

Пример. Пусть T6 = "Кем выпускается препарат "Зиннат"?".
Тогда Semrepr = *Вопрос (x1, Ситуация (e1, выпуск1 * (Время, #сейчас#) (Агент1, x1)(Продукция1, нек препарат1 * (Название, 'Зиннат') : x2)))* .

Пример. Пусть T7 = "Откуда и для кого поступил трехтонный алюминиевый контейнер?". Тогда Semrepr = *Вопрос ((x1 ∧ x2) , Ситуация (e1, поступление2 * (Время, нек мом * (Раньше, #сейчас#) : t1) (Место1, x1) (Адресат, x2) (Объект1, нек контейнер * (Вес, 3/ тонна))(Материал, алюминий) : x3))))* .

Пример. Пусть T8 = "Сколько человек участвовало в создании статистического сборника?". Тогда Semrepr = *Вопрос(x1, ((x1≡ Колич(S1)) ∧ Кач-состав (S1, чел) ∧ Описание(произв чел* (Элемент, S1) : y1, Ситуация(e1, участие1* (Время, нек мом * (Раньше, #сейчас#) : t1) (Агент1, y1)(Вид-деятельности, создание1* (Продукт1, нек сборник1 * (Область1, статистика) : x2)))))* .

Пример. Пусть T9 = "Сколько раз Иван Михайлович Семёнов летал в Мексику?". Тогда Semrepr = *Вопрос(x1, ((x1≡ Колич(S1)) ∧ Кач-состав (S1, сит) ∧ Описание(произв сит * (Элемент, S1) : e1, Ситуация (e1, полёт * (Время, нек мом * (Раньше, #сейчас#) : t1)(Агент1, нек чел.*(Имя, 'Иван')(Отчество, 'Михайлович')(Фамилия, 'Семёнов'): x2)(Место2, нек страна* (Название, 'Мексика') : x3)))))* .

7.4. Обсуждение разработанного метода преобразования ЕЯ-текстов в семантические представления

Изложенный метод, базирующийся на построенной формальной модели лингвистической базы данных (ЛБД) и на введенном понятии матричного семантико-синтаксического представления (МССП), направлен на непосредственное установление смысловых отношений между элементарными

значащими единицами входного текста, отражая эти отношения посредством МССП, и на последующее построение семантического представления (СП) текста, являющегося выражением некоторого СК-языка (К-представлением). Рассматриваемые тексты могут выражать высказывания (сообщения), команды, специальные вопросы (т.е. вопросы с вопросительными словами), общие вопросы (т.е. вопросы с ответом «Да»/ «Нет»). Тексты могут, в частности, включать причастные обороты и придаточные определительные предложения.

Метод позволяет устанавливать возможные смысловые отношения, в частности, в сочетаниях «Глагол + Предлог + Существительное», «Глагол + Существительное», «Существительное1 + Предлог + Существительное2», «Число + Существительное», «Прилагательное + Существительное», «Существительное1 + Существительное2», «Причастие + Существительное», «Причастие + Предлог + Существительное», «Вопросительно-относительное местоимение или местоименное наречие + Глагол», «Предлог + Вопросительно--относительное местоимение + Глагол».

Работоспособность изложенного метода реализации преобразования «Текст→СП» доказана созданием на его основе сложного структурированного алгоритма семантико-семантического анализа ЕЯ-текстов (см. главы 8, 9) и серии семантико-семантических анализаторов вопросов, команд и сообщений в системах программирования TurboPascal 7.0., C, C++, Delphi 4.0, 5.0, PHP.

Предложенный метод намечает принципиально новый подход к семантико-синтаксическому анализу ЕЯ-текстов.

Метод явно учитывает многозначность слов, что чрезвычайно важно для приложений и является его существенным преимуществом.

Важная особенность метода заключается в том, что он не предусматривает использования синтаксического уровня представления текста (как результата выполнения синтаксического анализа), в то время

как синтаксический уровень представления используется в течение нескольких десятилетий как в нашей стране, так и за рубежом.

Характер данных, описываемых формальной моделью ЛБД, и направленность предложенного метода на непосредственное выявление смысловых отношений между элементами текста с целью построения его СП позволяют провести некоторые параллели между разработанным методом и идеями компьютерной семантики русского языка.

Например, в статье (Тузов 2001) одно из возможных значений предлога “к” (значение “Куда”, т.е. значение “Направление движения”) представлено выражением *@ Куда К ($ 12 ~ @ Дат острову)* .

То же значение предлога “к” может быть отображено предложным фреймом вида *(j, ‘к ‘ , соб , простр.об , 4 , Куда, ‘к острову’)* из словаря *Frp* (см. параграф 6.7), являющегося компонентом лингвистического базиса Lingb. В последнем выражении *j* является порядковым номером фрейма, *соб* – сорт “событие”, *простр.об* – сорт “пространственный объект”, 4 - код винительного падежа, *Куда* – обозначение смыслового отношения “Направление движения”, *'к острову’* – пример реализации этого отношения.

Таким образом, в этом случае мы видим использование, по существу, одной и той же структуры данных. Однако необходимо отметить, что, в отличие от содержания данной главы и главы 7, в публикациях по компьютерной семантике русского языка не построена формальная модель ЛБД, не намечены контуры такой модели и не предлагается структура данных для представления промежуточных результатов семантико-синтаксического анализа ЕЯ-текстов.

Использование аппарата СК-языков для построения СП входных текстов ЛП позволило преодолеть трудности принципиального характера, касающиеся отображения содержания команд, а также вопросов нескольких видов: с вопросительными словами “какие”, “каких” и т.д., со словом “сколько”, относящимся к количеству предметов, и с ответом “Да /Нет”.

Важное преимущество изложенного нового подхода к разработке алгоритмов ССА заключается в создании предпосылок для облегчения подготовки специалистов в области лингвистических информационных технологий. Предложенный подход направлен на непосредственный поиск смысловых отношений между участниками ситуаций, и эти смысловые отношения понятны специалистам из рассматриваемой конкретной области (при этом предметная область может меняться). Как следствие, разработанный подход не требует овладения обширной лингвистической терминологией. Для понимания метода достаточно знакомства с базовыми математическими понятиями (множество, последовательность, цепочка, *n*-арное отношение, функция) и рядом понятий из курса русского языка по программе средней школы .