В.А. Фомичев

ФОРМАЛИЗАЦИЯ ПРОЕКТИРОВАНИЯ ЛИНГВИСТИЧЕСКИХ ПРОЦЕССОРОВ

МАКС ПРЕСС

МОСКВА 2005

В монографии описывается апробированная на практике новая система взаимосвязанных формальных моделей и алгоритмов, предназначенных для проектирования лингвистических процессоров (компьютерных систем, осуществляющих смысловую обработку письменных текстов или устной речи на естественном языке) в произвольных предметных областях. Значительное внимание уделяется изложению оригинального теоретического подхода к математическому описанию смысловой структуры не только предложений, но и сложных связных текстов (или дискурсов), относящихся к деловой прозе: текстов по медицине, экономике, юриспруденции и т.д. Анализируются возможности использования этого подхода в теории многоагентных систем, для разработки логико-информационных основ электронной коммерции и для устранения языкового барьера между пользователями сети Интернет из разных стран.

Значительная часть материалов монографии была опубликована в научных журналах “Информационные технологии”, “Качество и ИПИ (CALS)- технологии”, “Качество. Инновации. Образование”, “Informatica” (Словения), “Cybernetica” (Бельгия) и трудах международных научных конференций и симпозиумов, проходивших в России, Австрии, Великобритании, Германии, Дании, Нидерландах, Словении, Франции.

Книга не имеет аналогов в мировой научной литературе и будет полезна как опытным специалистам в области прикладных интеллектуальных систем или математической лингвистике, так и студентам и начинающим ученым.

The monograph describes a new system of interrelated formal models and algorithms tested in practice and destined for designing linguistic processors, or natural language processing systems (computer systems fulfilling the conceptual processing of written texts or oral speech in natural language) in arbitrary application domains. A considerable attention is drawn to setting forth an original theoretical approach to representing in a mathematical way the structured meanings (or conceptual structure, semantic structure) of not only separate sentences but also of complicated narrative texts (or discourses) pertaining to medicine, economy, law, and other fields of professional activity. The possibilities of using this approach in the multi-agent theory, for the elaboration of logical-informational foundations of electronic commerce (e-commerce), and for the elimination of the language barrier between the Internet users from various countries are analysed.

A considerable part of the stated materials was published in the scientific journals in Russian “Informational Technologies”, “Quality and IPI (CALS)-Technologies”, “Qulaity. Innovations. Education”, in the international scientific journals “Informatica” (Slovenia), “Cybernetica” (Belgium), and in the proceedings of the international scientific conferences and symposia which were held in Russia, Austria, Denmark, France, Germany, Slovenia, The Netherlands, and United Kingdom.

The monograph has no analogues in the world scientific literature and will be of use both to experienced specialists in the field of applied intelligent systems or mathematical linguistics and to the students and young scientists.

Моей семье:

Ольге Святославовне

Фомичевой,

Людмиле Дмитриевне

Удаловой,

Дмитрию Владимировичу

Фомичеву

посвящается

ОГЛАВЛЕНИЕ

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ

Всегда практика должна быть воздвигнута на
хорошей теории, ворота которой - перспектива
Леонардо да Винчи

В преподавании такой быстро развивающейся области,
какой является наука о вычислительных процессах,
правильный педагогический принцип состоит в том,
чтобы больше внимания уделять идеям, а не техни-
ческим подробностям реализации
А. Ахо, Дж. Ульман

За последние два десятилетия научно-техническое направление "искусственный интеллект" получило значительное развитие и нашло целый ряд успешных применений. Основная часть информации хранится и передается людьми с помощью естественного языка (ЕЯ), т.е. совокупности русского, английского, японского и других языков. Один из главных подклассов компьютерных систем с элементами искусственного интеллекта (СИИ) составляют программы, понимающие ЕЯ или синтезирующие выражения ЕЯ по некоторым внутренним представлениям. Такие программы называются системами обработки естественного языка (в англоязычной научной литературе: natural language processing systems), или лингвистическими процессорами (ЛП). Технологии, предусматривающие использование ЛП для обработки информации, составляют основной подкласс лингвистических информационных технологий (ЛИТ).

Другие виды современных ЛИТ связаны с разработкой и применением языков общения компьютерных интеллектуальных агентов (КИА) в многоагентных системах, языков построения протоколов переговоров, проводимых КИА в области электронной коммерции, и языков формирования контрактов, заключаемых КИА в ходе таких переговоров, а также семантически-структурированных языков нового поколения для представления информации во Всемирной Паутине (the World Wide Web, или WWW).

Несколько неформальных понятий, являющихся базовыми для теории смысловой обработки компьютером естественного языка, многократно используются в этой книге: семантика естественного языка, связный текст (или дискурс), структурированное значение выражения на ЕЯ, семантическое представление ЕЯ-выражения и алгоритм семантико-синтаксического анализа.

Под семантикой ЕЯ будем понимать совокупность закономерностей передачи информации средствами ЕЯ. Связным текстом (или дискурсом) называется последовательность взаимосвязанных по смыслу выражений на ЕЯ.

Если Т – некоторое выражение на ЕЯ (словосочетание, предложение, дискурс), то структурированным значением выражения Т является информационая структура, строящаяся мозгом человека, владеющего данным подязыком ЕЯ (русским, английским или другим), независимо от контекста, в котором услышано или прочитано выражение Т, т.е. строящаяся на основе только знаний о значениях элементарных лексических единиц и правил их комбинирования в данном языке.

Под семантическим представлением (СП) ЕЯ-выражения Т понимается формальная структура, являющаяся либо образом структурированного значения этого выражения, либо отражением смысла (или содержания) данного выражения в определенном контексте - в конкретной ситуации диалога, в контексте знаний о мире или в контексте предшествующей части дискурса.

Таким образом, СП ЕЯ-выражения Т является формальной структурой, первичными элементами которой являются, в частностия, обозначения понятий, конкретных объектов, множеств объектов, событий, имена функций и отношений, логические связки, обозначения чисел и цветов, а также обозначения смысловых отношений между значениями фрагментов текста или между объектами рассматриваемой предметной области.

СП текстов могут являться, например, строками и размеченными ориентированными графами (семантическими сетями).

Алгоритм семантико-синтаксического анализа строит по тексту на ЕЯ его СП, используя для этого знания о морфологии и синтаксисе подъязыка ЕЯ (русского, английского и др.), информацию о взаимосвязях лексических единиц с единицами семантического уровня и знания о мире. Семантическое

представление текста, построенное таким алгоритмом, интерпретируется прикладной интеллектуальной системой в зависимости от ее назначения, например, как задание на поиск ответа на вопрос, команда на выполнение физического действия автономным интеллектуальным роботом, фрагмент знаний о мире, предназначенный для пополнения базы знаний и т.д.

Научные результаты, изложенные в данной монографии, были получены автором в ходе цикла исследований, начатого более двадцати лет назад. Выбор направления исследований был реакцией на почти полное отсутствие в то время эффективных математических средств и методов проектирования ЛП.

Результаты данной монографии дают не только продвижение вперед, но и *качественный скачок* в области разработки формальных средств и методов проектирования алгоритмов семантико-синтаксического анализа ЕЯ-текстов. Этот качественный скачок обусловлен следующими основными факторами:

1. Разработчики ЛП получили систему правил (причем компактную, состоящую всего из 10 основных правил), позволяющих, по гипотезе автора, строить семантические представления произвольных текстов деловой прозы, т.е. текстов по экономике, технике, медицине, юриспруденции и т.д. Это означает, что эффективные процедуры построения СП ЕЯ-текстов и процедуры обработки СП ЕЯ-текстов (в контексте содержания предшествующей части текста или диалога, в рамках знаний о предметной области и т.д.) можно будет использовать в разных предметных областях и развивать возможности этих процедур при возникновении новых задач.
2. Построена формальная модель лингвистической базы данных, содержащей такие сведения о лексических единицах и их взаимосвязях с информационными единицами, которые достаточны для семантико-синтаксического анализа интересных для приложений подъязыков русского языка.
3. Разработан практически полезный сложный структурированный алгоритм семантико-синтаксического анализа, который описывается не средствами какой-либо системы программирования, а полностью с помощью

предложенной системы формальных понятий, что делает этот алгоритм независимым от программной реализации и предметной области.

СОДЕРЖАНИЕ КНИГИ

В главе 1 дается краткий обзор областей применения лингвистических процессоров, а также анализируются потребности расширения запаса эффективных формальных средств и методов для проектирования ЛП и разработки ЛИТ в области многоагентных систем и электронной коммерции.

В главе 2 описывается математическая модель, перечисляющая первичные единицы концептуального уровня, используемые ЛП, а также описывающая информацию, связанную с такими единицами и необходимую для соединения этих единиц в составные единицы, отображающие структурированные значения (СЗ) сколь угодно сложных ЕЯ-текстов.

В главе 3 (в развитие результатов главы 2) построена математическая модель для описания СЗ предложений и сложных связных текстов (дискурсов) на естественном языке (в частности, на русском, английском, немецком, французском языках). Модель представляет собою определение нового класса формальных языков, названных стандартными концептуальными языками (стандартными К-языками, СК-языками), и может рассматриваться как формальная грамматика нового вида. Сущность этой модели в том, что она задает 10 операций на концептуальных структурах, с помощью которых за конечное число шагов можно построить семантическое представление предложения или дискурса из чрезвычайно широкого подъязыка деловой прозы.

Проведено математическое исследование формальных объектов, задаваемых этой моделью – выражений СК-языков. В частности, доказана однозначность структурного анализа таких выражений.

Глава 4 посвящена исследованию выразительных возможностей класса СК-языков. Показано, что выражения СК-языков удобно использовать для: (а) построения СП предложений (выражающих высказывания, вопросы, команды) и сложных дискурсов на русском языке, (б) построения составных целей, (в)

представления знаний о мире, в том числе для построения формальных определений понятий и объектно-ориентированных модулей знаний..

Проведено сравнение выразительных возможностей СК-языков с выразительными возможностями других, наиболее часто используемых подходов к формальному представлению значений (смысловой структуры) ЕЯ-текстов: теории представления дискурсов, теории концептуальных графов, эпизодической логики, теории расширенных семантических сетей, теории неоднородных семантических сетей и компьютерной семантики русского языка. Показано, что выразительные возможности СК-языков значительно превосходят возможности перечисленных подходов и, в то же время, аппарат СК-языков позволяет моделировать механизмы представления информации, характерные для каждого из указанных подходов.

В главе 5 исследуются возможности использования аппарата СК-языков для решения ряда актуальных проблем информатики: разработки языков представления содержания посланий компьютерных интеллектуальных агентов, в частности, языков, предназначенных для формирования контрактов и протоколов переговоров в области электронной коммерции, создания семантического сетевого языка нового поколения, построения онтологий предметных областей, разработки новых языков представления знаний для решения информационно-сложных задач, проектирования интеллектуальных информационно-поисковых и вопросо-ответных Интернет-систем нового поколения.

В главе 6 вводится формальное понятие лингвистического базиса, которое интерпретируется как описание структуры лингвистической базы данных (ЛБД), используемой алгоритмом семантико-синтаксического анализа ЕЯ-текстов. ЛБД, структура которых отображается построенной моделью, позволяют 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 Наречие + Глагол», «Предлог + Вопросительно- относительное местоимение + Глагол».

В главе 7 излагается новый метод преобразования ЕЯ-текстов в их семантические представления. Метод предусматривает использование предложенного автором матричного семантико-синтаксического представления (МССП) входного текста как промежуточного представления при переходе от ЕЯ-текста к СП текста, являющемуся выражением некоторого СК-языка (т.е. К-представлением текста). При этом не используется традиционное синтаксическое представление текста. Тексты могут быть, в частности, вопросами, сообщениями (описаниями фактов, ситуаций) или командами.

В главах 8 и 9 разработан сложный структурированный алгоритм семантико-синтак-сического анализа текстов из представляющих практический интерес подъязыков естественного (русского) языка (алгоритм SemSyn). Этот алгоритм, базирующийся на построенной в главе 6 формальной модели ЛБД и на введенном в главе 7 понятии МССП текста, устанавливает смысловые отношения между элементарными значащими единицами входного текста, отражая эти отношения посредством МССП, а затем строит СП текста, являющееся выражением некоторого СК-языка (К-представлением). Входные ЕЯ-тексты могут выражать высказывания (сообщения), команды, специальные вопросы (т.е. вопросы с вопросительными словами), общие вопросы (т.е. вопросы с ответом «Да»/ «Нет»)и могут, в частности, включать причастные обороты и придаточные определительные предложения. Алгоритм SemSyn позволяет устанавливать возможные смысловые отношения, в частности, в сочетаниях перечисленных выше видов.

В заключении к данной монографии делается вывод о том, что совокупность научных результатов, изложенных в главах 1 - 4, 6 - 9, и часть научных результатов главы 5 образуют новую теорию проектирования семантико-синтаксических анализаторов естественно-языковых текстов с использованием формальных средств представления входных, промежуточных и выходных данных; эта теория может быть названа теорией К-представлений.

Приложение содержит доказательства двух лемм и базирующегося на них доказательства одного из утверждений из главы 3. Нумерация утверждений сквозная внутри каждой главы (Утверждение 3.1, Утверждение 3.2 и т.д.).

В основе большей части содержания данной монографии лежат циклы лекций, читавшиеся автором с 1996 г. студентам Российского государственного технологического университета им. К.Э Циолковского – “МАТИ” по дисциплинам “Теоретические основы лингвистических информационных технологий”, “Математическая лингвистика”, “Проектирование лингвистических процессоров” и студентам Московского государственного института электроники и математики (технического университета) по дисциплинам “Лингвистические информационные технологии”, “Проектирование лингвистических процессоров” и “Глобальные информационные сети и дистанционное обучение”.

БЛАГОДАРНОСТИ

Я благодарен профессору, д.т.н., зав. кафедрой “Программное обеспечение вычислительных машин” Российского государственного социального университета Ю.П. Кораблину, профессору МАТИ Г.С. Плесневичу, профессорам МИЭМ Л.С. Воскову и А.К. Зыкову за обсуждение многих разделов данной монографии и полезные замечания, а также профессору, д.т.н., заслуженному деятелю науки и техники РСФСР , зав. кафедрой “Системы автоматического управления” МГТУ им. Н.Э. Баумана К.А. Пупкову за поддержку первых, самых трудных шагов исследования, результаты которого представлены в данной монографии.

С 1990-х годов положение науки в нашей стране, к сожалению, остается таким, что появление этой книги было бы невозможно без огромной поддержки, внимания, терпения моей жены - Ольги Святославовны Фомичевой, и мамы Ольги Святославовны - Людмилы Дмитриевны Удаловой.

Благодаря помощи моего сына Димы, выпускника факультета вычислительной математики и кибернетики МГУ им. М.В. Ломоносова, в освоении нескольких компьютерных технологий были подготовлены к печати многие работы, послужившие основой для этой монографии.

Я признателен директору издательства МАКС Пресс, Алле Николаевне Матвеевой, за предложение подготовить и издать эту книгу.

Большая помощь в подготовке в электронном виде материалов, послуживших основой для книги, была оказана многими студентами кафедры “Информационные технологии” МАТИ, особенно Я.В. Ахромовым, и студентами кафедры “Математическое и программное обеспечение систем обработки информации и управления” Московского государственного института электроники и математики (технического университета).

Глава 1

ФОРМАЛИЗАЦИЯ СЕМАНТИКИ ЕСТЕСТВЕННОГО ЯЗЫКА И ПОТРЕБНОСТИ ПРОЕКТИРОВАНИЯ ЛИНГВИСТИЧЕСКИХ ПРОЦЕССОРОВ

1.1. Области применения лингвистических процесоров

Прогресс, достигнутый за последние два десятилетия в области проектирования ЛП, выразился в появлении широкого спектра областей применения ЛП. Такими областями, в частности, являются: машинный перевод

письменных текстов (исторически первая область использования ЛП) и устной речи; естественно-языковые интерфейсы (ЕЯ-интерфейсы) прикладных интеллектуальных систем: экспертных систем, расчетно-логических систем, автономных интеллектуальных роботов; синтез текстов, представляющих рекомендации пользователю экспертной системы (медицинской диагностики, технической диагностики и др.) в естественно-языковой форме; проектирование концептуальных схем баз данных посредством преобразования ЕЯ-спецификаций предметной области в концептуальную схему базы данных; автоматизированное проектирование технических объектов (например, электронных блоков) с помощью преобразования ЕЯ-спецификации проектируемого объекта в формальную спецификацию и затем – в проектную документацию технического объекта.

Развитие исследований в области конструирования ЛП привело к появлению новых теоретических и практических задач.

Государственными и коммерческими организациями накоплены большие запасы информационных ресурсов, содержащих знания о предметных областях. Для повышения эффективности работы сотрудников с накопленными знаниями крупные компании в мире разрабатывают или уже разработали и используют системы управления знаниями. По имеющимся в литературе оценкам, более 70% ресурсов, накопленных в различных организациях, носит неструктурированный характер и образуется электронными текстовыми документами. Поэтому, по мнению ряда авторов, повышению эффективности работы сотрудников различных организаций с накопленными информационными ресурсами будет способствовать разработка интеллектуальных поисковых систем с ЕЯ-интерфейсами, способных осуществлять смысловой анализ естественно-языковых полей разнообразных используемых электронных документов и, как следствие, давать ссылки на документы, интересующие пользователя, или формулировать ответы на поставленные пользователем вопросы (Попов 2001, 2002; Королев 2003; Арлазаров, Емельянов 2003, 2004; Pohl 2003).

Создание таких интеллектуальных поисковых систем с ЕЯ-интерфейсами, и особенно Интернет-систем, представляется весьма актуальным направлением развития исследований по разработке CALS (ИПИ)-технологий

Непрерывная информационная поддержка жизненного цикла сложного изделия предполагает совместное использование субъектами виртуального предприятия (одной из современных форм реализации CALS (ИПИ)-технологий) единой базы знаний о рассматриваемых предметных областях (возможно, распределенной и с некоторыми ограничениями на конфигурацию базы знаний, доступную определенному субъекту виртуального предприятия) и эффективный обмен информацией между субъектами виртуального предприятия. В этой связи ЕЯ-интерфейсы обещают упростить и, как следствие, увеличить эффективность взаимодействия непрограммирующих специалистов с базами данных и базами знаний

Другой острой проблемой теории СИИ является автоматизация формирования баз знаний (БЗ) СИИ. Основная часть знаний, накопленных человечеством, хранится в виде естественно-языковых текстов (ЕЯ-текстов). Поэтому в последние годы реализован или реализуется ряд проектов, направленных на автоматическое извлечение знаний из ЕЯ-текстов. Значительное внимание в Германии, США, Японии и некоторых других странах уделяется проблеме автоматизации извлечения знаний из биологических и медицинских документов (отчетов об исследованиях, статей в научных журналах и т.д.). Проекты по этой проблеме составляют важную часть нового направления в информатике, получившего название *биоинформатика.*

Однако построенные системы извлечения знаний из ЕЯ-текстов обладают весьма узкими способностями понимания ЕЯ-текстов, особенно связных текстов (дискурсов), т.е. последовательностей взаимосвязанных по смыслу фраз на ЕЯ. Это выражается в использовании разнообразных узкоспециализированных шаблонов для извлечения знаний. Центральной причиной этого положения является недостаточная проработанность вопросов формального описания закономерностей передачи информации средствами ЕЯ, т.е. вопросов формализации семантики ЕЯ.

Благодаря бурному прогрессу компьютерной сети Всемирная Паутина (the World Wide Web, WWW, W3) пользователи сети во всем мире получили быстрый доступ к огромному количеству ЕЯ-текстов, относящихся к различным областям деятельности человека. С середины 1990-х годов специалисты в самых разных предметных областях работают не только с публикациями и базами данных (БД) своих организаций, но и стремятся использовать информационные ресурсы Паутины. Поэтому чрезвычайно актуальна задача организации взаимодействия на ограниченном естественном языке из различных предметных областей с огромным объемом накопленных информационных ресурсов Всемирной Паутины (Попов 2002; Хорошевский 2002).

ЕЯ-интерфейсы для взаимодействия с информационными ресурсами Паутины необходимы не только специалистам для решения профессиональных задач, но и конечным пользователям, перед которыми стоят задачи получения медицинской или юридической информации, расширения культурного кругозора, получения дополнительного профессионального образования и т.д.

В феврале 2001 г. консорциум сети Всемирная Паутина, обозначаемый в большинстве документов сокращением W3C (the World Wide Web Consortium), официально объявил о широком развертывании исследований по преобразованию существующей сети в Семантическую Всемирную Паутину (Semantic Web). Один из наиболее важных аспектов реализации этого крупномасштабного проекта заключается в том, что компьютерные интеллектуальные агенты (КИА) смогут анализировать информацию, представленную на Веб-сайтах, взаимодействуя между собой. Часть КИА сможет выполнять смысловой анализ естественно-языковых компонентов электронных документов, представленных в Веб-сайтах. Это даст возможность конечным пользователям осуществлять поиск информации в Паутине не по ключевым словам, а по смыслу, с помощью КИА (Semantic Web 2001).

Важные дополнительные возможности для пользователя предоставят речевые браузеры: они позволят использовать телефоны (в том числе мобильные) для взаимодействия с Семантической Паутиной на ЕЯ (Voice 2001).

Прогресс в разработке компьютеров, ЛП и средств телекоммуникации привел в 1990-е годы к реализации в ряде стран проектов создания электронных

библиотек (ЭлБ), называемых в англоязычной литературе цифровыми библиотеками. В нашей стране важными импульсами к развертыванию научно-технической программы в этом направлении стали Российско-американский семинар, проходивший в 1998 г, и первая национальная конференция по электронным библиотекам с участием ученых из Германии и США, состоявшаяся в 1999 г. в Санкт-Петербурге. В итоговых материалах этой конференции, в частности, отмечается, что одной из центральных научных задач, связанных с созданием ЭлБ, является автоматизация семантического анализа ЕЯ-текстов с целью смыслового поиска информационных источников.

Развитие гражданского общества в нашей стране существенно зависит от степени доступности государственных информационных ресурсов. Обеспечение такой доступности является одной из центральных задач федеральной целевой программы “Электронная Россия (2002 – 2010 годы)“. Огромную роль в обеспечении доступа общественности к государственным информационным ресурсам должны сыграть ЭлБ. Для обеспечения подлинной широты доступа пользователей ЭлБ к информационным ресурсам необходимы интеллектуальные поисковые системы с ЕЯ-интерфейсами, способные отыскивать информационные источники или находить ответы на вопросы конечных пользователей на основе осуществления смыслового анализа (а) запроса пользователя, (б) естественно-языковых полей разнообразных хранящихся электронных документов и сравнения содержания запроса пользователя с содержанием анализируемых текстовых полей электронных документов.

В свете перечисленных и ряда других направлений применения ЛП, разработка теории и методов компьютерного понимания ЕЯ-текстов и извлечения знаний из ЕЯ-текстов является важным направлением развития теории интеллектуальных компьютерных систем (Арлазаров, Журавлев, Ларичев и др. 1998). Этой проблеме было уделено значительное внимание на Научной сессии Отделения информационных технологий и вычислительных систем РАН, состоявшейся в мае 2003 года.

1.2. Значение формальных методов для разработки лингвистических информационных технологий

Накопленный опыт исследований по созданию ЛП показал, что огромное влияние на проектирование анализаторов ЕЯ-текстов оказывают используемые методы формального отображения содержания (или смысла) текстов, а также методы формального представления промежуточных результатов смыслового анализа текстов. Особую актуальность в 1990-е годы приобрела проблема формального представления содержания связных текстов (или дискурсов).

Во-первых, основной объем информации в текстовых БД и сети Интернет представлен дискурсами. Во-вторых, сформулированная Э.В. Поповым современная концепция разработки систем общения с БД на ограниченном естественном языке (ОЕЯ) предполагает, что на вход системы поступают не только предложения, но и дискурсы (Попов 2002). В-третьих, можно согласиться с высказанной Э.В. Поповым гипотезой о том, что повышению эффективности общения на ОЕЯ с большими БД будет способствовать реализация таких систем общения, когда активную роль в диалоге будет играть не только конечный пользователь, но и компьютер, располагающий моделью базы знаний, причем инициатива будет на протяжении диалога неоднократно переходить от одного участника общения к другому. Последовательность выражений на ОЕЯ (с указанием авторов выражений), сформированных участниками общения, образует дискурс.

Можно выделить несколько наиболее важных аспектов проблемы формального представления содержания (или смысла) ЕЯ-текстов в компьютерных системах.

Идея использования в системах машинного перевода искусственного языка-посредника для представления смысла ЕЯ-текстов была высказана еще в 1960-м году А.К. Жолковским, Н.Н. Леонтьевой и Ю.С. Мартемьяновым. В 1960-е – 1970-е годы эта идея получила значительное развитие в работах А.К. Жолковского и И.А. Мельчука по лингвистической модели “Смысл – Текст” (Жолковский, Мельчук 1969; . Мельчук 1974). В 1970-е годы усилению внимания к идее семантического языка-посредника способствовала теория

смысловой зависимости в ЕЯ Р. Шенка, нашедшая применение в нескольких экспериментальных системах компьютерной обработки ЕЯ (Schank 1972; . Schank и др. 1975).

Использование языка-посредника для представления содержания (смысла) ЕЯ-текстов позволяет перейти от неформализованного объекта, каким является ЕЯ-текст, к формальной структуре, что открывает возможности обработки этой структуры различными процедурами – “семантическими экспертами” в рамках базы знаний, представленных записями на формальном языке (языке представления знаний). На протяжении 1980-х – 2000-х годов в проектировании ЛП наиболее часто использовались языки-посредники, предоставляемые теорией семантических сетей, теорией фреймов, теорией концептуальных графов и эпизодической логикой. В нашей стране использовался также язык-посредник, разработанный в рамках компьютерной семантики русского языка, расширенные семантические сети, неоднородные семантические сети (см. параграф 4.15), стандартные К-языки, предложенные автором данной работы, и некоторые другие подходы.

В середине 1990-х годов возникла новая проблема, усилившая внимание исследователей к проблеме разработки языка-посредника для отображения содержания ЕЯ-текстов. С целью устранения языкового барьера между пользователями сети Интернет из разных стран мира, Х.Учида и М. Жу (Япония) предложили новый язык-посредник, использующий слова английского языка для обозначения информационных единиц и несколько специальных символов. Этот язык, названный универсальным сетевым языком (UNL, the Universal Networking Language), базируется на идее отображения содержания фраз с помошью бинарных отношений. С конца 1990-х годов ООН финансируется комплексный проект, направленный на разработку системы ЛП, преобразующих фразы на различных естественных языках в выражения языка UNL, а также преобразующих выражения языка UNL в предложения на различных естественных языках. Координатором проекта является Институт передовых исследований ООН Токийского университета. В настоящее время в проекте разрабатываются ЛП для шести официальных языков ООН (английского, арабского, испанского, китайского, русского и французского), а

также для хинди, индонезийского, итальянского, японского, латышского, немецкого, монгольского, португальского, суахили и тайского языков (Uchida, Zhu, Della Senta 1999; Uchida, Zhu 2001; Zhu, Uchida 2002).

Проблема создания широко применимых методов формального описания содержания (смысла) предложений и дискурсов (другими словами, описания структурированных значений ЕЯ-текстов) тесно соприкасается с потребностями развития таких бурно развивающихся направлений информатики, как многоагентные системы (МАС) и электронная коммерция. Взаимодействие компьютерных интеллектуальных агентов (КИА) осуществляется через обмен посланиями (messages), которые могут выражать сообщения, вопросы и команды. Для формирования таких посланий разрабатываются специальные языки общения интеллектуальных агентов (Agent Communication Languages, или ACL). Для координации деятельности исследовательских центров разных стран по разработке стандартных инструментальных средств в области МАС в 1996 г. образован международный Фонд интеллектуальных физических агентов (The Foundation for Intelligent Physical Agents, или FIPA), штаб-квартира которого находится в Женеве. В 1997 - 2000 годах в рамках этого фонда был разработан стандарт языка общения КИА, который в дальнейшем будет называться FIPA ACL. Часть этого языка, предназначенная для представления содержания посланий (в отличие от внешней информации - об отправителе, получателе и т.д.), названа семантическим языком (FIPA Semantic Language, или FIPA SL). Фондом поставлена задача разработки библиотеки языков представления содержания посланий КИА (Content Languages), совместимых с этим языком и охватывающих весь спектр применений МАС.

Многоагентные системы рассматриваются как ключевая технология для реализации электронной коммерции. Следовательно, выразительные возможности языка общения КИА должны быть достаточными для того, чтобы представлять содержание произвольных коммерческих переговоров и контрактов, заключенных в результате этих переговоров. Поэтому формальные языки для представления содержания коммерческих переговоров и контрактов являются предметами исследования в новых научных направлениях в области

МАС, называемых *электронными переговорами* (e-negotiations) и *электронным заключением контрактов* (electronic contracting).

Между тем, выразительные возможности семантического языка FIPA SL довольно далеки от того, чтобы быть удобными для решения этой задачи. В связи с этим актуальна задача создания методов разработки более совершенных формальных языков - таких, которые были бы удобны для представления содержания любых посланий КИА, в том числе и для представления содержания произвольных коммерческих переговоров и контрактов.

Проблема разработки формальных языков-посредников для отображения содержания (или смысла) ЕЯ-текстов (другими словами, языков семантических представлений, или семантических языков) исследуется специалистами разных стран в течение более трех десятилетий. В нашей стране ряд аспектов этой проблемы в различные периоды изучались Ю.Д. Апресяном, И.М. Богуславским, В.М. Брябриным, Б.Ю. Городецким, А.К. Жолковским, А.П. Ершовым, Ю.И. Клыковым, О.С. Кулагиной, Е.С. Кузиным, Л.Т. Кузиным, И.П. Кузнецовым, Д.Г. Лахути, Н.Н. Леонтьевой, Л.И. Литвинцевой, Ю.Я. Любарским, М.Г. Мальковским, А.Г. Мацкевичем, И.А. Мельчуком, Л.И. Микуличем, А.С. Нариньяни, Г.С. Осиповым, Г.С. Плесневичем, Э.В. Поповым, Д.А. Поспеловым, В.Ш. Рубашкиным, В.А. Тузовым, З.М. Шаляпиной, Г.С. Цейтиным, Л.Л. Цинманом и другими учеными.

За рубежом наибольший вклад в разработку методов математического описания содержания (смысла) ЕЯ-текстов внесли Р. Монтегю (грамматики Монтегю), Дж. Барвайз и Р. Купер (теория обобщенных кванторов, ситуационная теория), М. Кресвелл (теория структурированных значений предложений), Й. Гронендейк и М. Стокхоф (динамические грамматики Монтегю, динамическая предикатная логика), Дж. Сова (теория концептуальных графов), Л. К. Шуберт и Ч.Х. Хуан (эпизодическая логика), Г. Камп и У. Рейль (теория представления дискурсов)

Несмотря на усилия, предпринимавшиеся в течение многих лет учеными разных стран, до последнего времени многие существенные аспекты проблемы формального описания содержания ЕЯ-текстов оставались мало изученными. Одна из основных причин этой ситуации заключается в том, что внимание

уделялось, главным образом, формализации смысловой структуры отдельных фраз, а не дискурсов. Кроме того, недостаточно изученной является проблема формального описания смысловой структуры отдельных фраз, обозначающих высказывания и включающих описания множеств и/или придаточные цели и/или слова “понятие”, “термин”, а также структуры фраз, выражающих команды и вопросы.

Наконец, сегодня ясно, что понимание ЕЯ-текста осуществляется в контексте системы знаний о мире и о целях интеллектуальных систем. Однако выразительные возможности большинства известных подходов к математическому описанию смысловой структуры ЕЯ-текстов (а именно, грамматик Монтегю, теории обобщенных кванторов, ситуационной теории, теории структурированных значений предложений, динамических грамматик Монтегю, динамической предикатной логики) недостаточны для построения теорий компьютерного понимания ЕЯ в контексте системы знаний о мире и о целях интеллектуальных систем. Например, исследования по дескриптивным логикам, выросшие из работ по терминологическим языкам представления знаний (ЯПЗ), показали полезность включения в состав ЯПЗ составных обозначений понятий. Однако перечисленные непосредственно выше подходы не предоставляют такой возможности.

Проблема автоматизации формирования баз знаний СИИ посредством извлечения информации из ЕЯ-текстов с помощью ЛП, проблема разработки семантического языка-посредника для устранения языкового барьера между пользователями сети Интернет и ряд других актульных научно-технических проблем требуют создания эффективных средств формального представления содержания произвольных ЕЯ-текстов, относящихся к *деловой прозе* (термин А.П. Ершова, ставший широко популярным в компьютерной лингвистике), т.е. ЕЯ-текстов, относящихся к юриспруденции, бизнесу, медицине, технике и т.д.

Между тем, перечисленные наиболее популярные подходы к формальному представлению содержания ЕЯ-текстов имеют ограниченную сферу применения. В частности, эти подходы не предоставляют адекватных формальных средств для представления содержания произвольных предложений с описаниями множеств или составными обозначениями понятий,

дискурсов со ссылками на смысл фраз и более крупных частей текстов, с обозначениями сложных целей, с косвенной речью.

Так, язык-посредник UNL ориентирован на представление содержания отдельных предложений, а не дискурсов. Кроме того, в языке UNL нет формальных средств описания множеств, средств формального различения описаний объектов и описаний понятий, квалифицирующих эти объекты, средств представления ссылок на смысл фраз и более крупных фрагментов дискурсов.

В связи с этим актуальна проблема разработки более мощных математических методов описания смысловой структуры реальных предложений и связных текстов, относящихся к юриспруденции, бизнесу, медицине, технике, экономике и т.д.

Наибольшие трудности при разработке ЛП связаны с выполнением преобразования “ЕЯ-текст ➔ Семантическое представление (СП) 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 структуры лингвистической базы данных (ЛБД), т.е. базы данных (БД) с морфологической и семантико-синтаксической информацией о лексических единицах, а также 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). Чрезвычайно существенно то, что полные описания информационного и программного обеспечения таких 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 специалистов в нашей стране в области

проектирования ЛП и сужает возможности принятия оптимальных проектных решений, приводит к дополнительным трудозатратам на разработку ЛП.

Учитывая сказанное, 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 большой трудоемкостью. В связи с этим в данной книге выдвигается 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 представления промежуточных и окончательных результатов семантико-синтаксического анализа ЕЯ-текстов по отношению к многообразию решаемых задач, предметных областей и программных сред (преемственность понимается как многократное, максимальное использование эффективных алгоритмов, реализуемых подсистемами ЛП, в разных проектах ЛП).

В данной работе предпринята попытка создания значительной части предпосылок для реализации этих двух принципов при проектировании ЛП.

1.3. Подходы к формализации семантики естественного языка, разработанные в конце 1960-х – первой половине 1980-х годов

Основные результаты в теории лингвистических процессоров (ЛП) были получены с конца 1960-х годов. В 1970-е годы было достигнуто значительное продвижение вперед в отношении принципов "понимания" компьютерной системой естественного языка (ЕЯ). В большой степени этому способствовали работы Т.Винограда (Winograd 1971), У.Вудса и Р.Каплана (Woods, Kaplan 1971), Й. Уилкса (Wilks 1973), Р. Шенка и его коллег Ч. Ригера, Н. Голдмана, Ч. Ризбека (Schank и др. 1975). Проекты, реализованные этими исследователями, показали, в частности, что (а) понимание ЕЯ-выражения осуществляется в рамках базы знаний (БЗ) о мире и (б) целесообразно использовать специальные формальные выражения для отображения смысла ЕЯ-выражений, причем выбор подхода к построению этих формальных выражений значительно влияет на процесс разработки ЛП.

Наиболее популярными подходами к построению таких формальных структур для отображения смысла ЕЯ-текстов в 1970-е годы являлись теория семантических сетей (ТСС), импульс к появлению которой был дан работами Куиллиана (Quillian 1968) и Р.Саймонса (Simmons 1973), а также теория смысловой зависимости в естественном языке (ТСЗЕЯ) Р. Шенка (Schank 1972; Schank и др. 1975). Как известно, семантические сети являются ориентированными графами со специальными метками вершин и ребер. Метки вершин обозначают понятия, реальные предметы, ситуации (в частности, события), числа, значения цветов и т.д., а метки ребер соответствуют смысловым отношениям между элементами текста и/или между понятиями. ТСЗЕЯ предложила способы построения диаграмм определенных видов для отображения смысла фраз и коротких связных текстов. Оба этих подхода не были математическими, но ТСС использовала математическое понятие

ориентированного графа для иллюстрации способов представления содержания простых фраз и текстов в виде размеченного графа.

В нашей стране в 1970-е и 1980-е годы в процессе развития теории семантических сетей И.П. Кузнецовым (Кузнецов 1976, 1978, 1986) возникла теория расширенных семантических сетей (см. параграф 4.15 данной книги).

Успехи 1970-х годов по реализации экспериментальных проектов понимания компьютером отдельных фраз на ЕЯ и текстов, состоящих из нескольких простых фраз, позволили выдвинуть в 1980-е годы перед исследователями следующие новые задачи: (а) переход от обработки простых фраз к обработке связных текстов (дискурсов), включающих пропуски слов в отдельных фразах (явление эллипсиса), ссылки на ранее упомянутые объекты ("для этого предприятия" и т.п.) и ссылки на смысл предыдущих фраз и более крупных частей текста ("об этом", "этот метод" и т.п.); (б) эффективный учет прагматики общения, т.е. интерпретация очередного ЕЯ-выражения в контексте всего диалога; (в) создание формальных, предметно-независимых методов проектирования ЛП с целью получения возможности широкого тиражирования эффективных проектных решений.

Вопрос о необходимости эффективных формальных методов для проектирования ЛП возник совершенно естественно. Дело в том, что опыт реализации в 1970-х годах экспериментальных проектов ЛП показал, что ЛП, обеспечивающие информационные потребности реального пользователя в той или иной предметной области, будут являться сложными программными комплексами. При разработке сложных технических систем в различных предметных областях широко используются математические методы. Например, для конструирования самолетов разработана и используется аэродинамика, а для проектирования кораблей и подводных лодок применяется гидромеханика.

Достаточно полное представление о том, какие формальные инструменты для изучения семантики ЕЯ были доступны к середине 1980-х годов разработчикам ЛП за рубежом, дают учебник (Thayse и др. 1988) по логическим методам в научном направлении Искусственный Интеллект (ИИ), учебник (Partee, ter Meulen. и Wall, 1990) по математической лингвистике и учебники (Grishman 1986; Gazdar, Mellish 1989) по компьютерной лингвистике. Анализ этих

источников и целого ряда других публикаций показывает, что в этот период запас формальных методов для изучения семантики ЕЯ был довольно бедным в отношении математического описания смысловой структуры реальных связных текстов (или дискурсов) на ЕЯ и соответствия между текстами и их семантическими представлениями (СП) с учетом базы знаний о мире.

Общепринято считать, что история современных подходов к формализации семантики ЕЯ начинается с фундаментальных работ американского логика Р. Монтегю (Montague 1970, 1974a, 1974b). Подход к формализации семантики ЕЯ, изложенный в этих работах, впоследствии был назван Грамматикой Монтегю. Рядом исследователей были предложены различные расширения Грамматики Монтегю; в частности, такие расширения рассматриваются в работах (Partee 1976; Thomason 1980). В 1980-х годах было предпринято несколько попыток использовать подход Монтегю в проектировании ЛП. Однако теория формализации семантики ЕЯ, получившая название Грамматики Монтегю, (а) не является универсальной, (б) недостаточно удобна для практики с точки зрения вычислительной эффективности, (в) изложена ее автором в трудно воспринимаемой форме, что тормозило прямое использование этой теории на практике. Поэтому подход Монтегю был переработан и дополнен (иногда существенно) с целью использования его в проектировании ЛП. Так, Клиффорд (Clifford 1983, 1988) разработал определение формального языка QE-III для представления содержания вопросов к историческим базам данных. Это было сделано на основе расширения Грамматики Монтегю. Хирст (Hirst 1988) разработал фреймоподобный семантический язык FRAIL, отойдя довольно далеко от Грамматики Монтегю. Язык FRAIL использовался в семантическом интерпретаторе ABSITY для представления результатов обработки ЕЯ-текстов с целью включения этих результатов в базу знаний. Джоуси (Jowsey 1987) предложил упрощенную версию Грамматики Монтегю для построения СП текстов в прикладной интеллектуальной системе, выполняющей рассуждения общего вида. Сембок и Райсберген (Sembok & van Rijsbergen, 1990) применили язык Джоуси, близкий к языку логики первого порядка, в экспериментальной информационно-поисковой системе.

Обобщенные грамматики фразовых структур (Gazdar, Klein, Pullum, Sag, 1985) также могут расматриваться как расширения Грамматики Монтегю, поскольку в этих грамматиках язык интенсиональной логики Монтегю используется для построения СП предложений.

Другими основными подходами к формальному изучению семантики ЕЯ в первой половине 1980-х годов были теория обобщенных кванторов (Barwise, Cooper 1981; Gaerdenfors 1987; Peres 1991), ситуационная семантика (Barwise, Perry 1983; Fenstad, Halvorsen и др. 1987; Cooper 1991), теория представления дискурсов (Kamp, 1981; Kamp, Reyle 1990). Все эти подходы имеют общую отправную точку – пионерские работы Р. Монтегю - и ряд взаимосвязей.

Анализ показывает, что глубокая связь с традициями математической логики является главной причиной очень большого разрыва между возможностями этих подходов и требованиями, предъявляемыми практикой проектирования ЛП. В частности, можно выделить следующие ограничения этих подходов к формальному изучению ЕЯ:

1. Неадекватность с точки зрения описания структурированных значений дискурсов на ЕЯ. В частности, отсутствие выразительных возможностей для описания семантической структуры дискурсов, содержащих ссылки на смысл фраз и более крупных частей текста (такие ссылки могут задаваться, например, словами и выражениями “поэтому”, “об этом”, “данный метод”, “’это распоряжение”, “поставленный вопрос”).

2. Идущая от логики внутренняя ориентация на рассмотрение выражений, представляющих высказывания. Между тем, еще существуют выражения, обозначающие вопросы, команды, цели, действия, пожелания, советы, обещания, назначения вещей, поэтому необходим формальный аппарат для описания смысловой структуры таких выражений. В этой связи следует отметить, что за рубежом первые шаги в этом направлении были сделаны теорией структурированных значений предложений (Cresswell 1985; Chierchia 1989), а в нашей стране первые шаги такого рода были сделаны в публикациях автора (Фомичев 1978а, 1981 а, б, 1983).

3. Игнорирование или недостаточно глубокое рассмотрение многих важных особенностей структуры выражений, обозначающих высказывания. В частности, можно отметить отсутствие адекватных формальных методов описания: (а) множеств,операций над множествами и отношений на множествах; (б) назначений вещей; (в) семантической структуры фраз, содержащих причастные обороты и придаточные определительные предложения; (г) стрктуры фраз, в которых логические связки “и“, “или“ соединяют не обозначения высказываний, а обозначения различных объектов, понятий, множеств или назначений вещей (“считывание или запись данных”, “прием и отправка груза” и т.п.); (д) структуры предложений со словами “понятие”, “термин”.

4. Структура данных, позволяющих поставить в соответствие выражению на ЕЯ одно или несколько возможных семантических представлений (СП) либо не моделировалась, либо моделировалась нереалистично с точки зрения разработки ЛП, способных анализировать тексты, относящися к науке, технике, экономике, медицине или юриспруденции.

5. Хорошо известно, что понимание ЕЯ-текста человеком может существенно зависеть от знаний этого человека (другими словами, реципиента текста) о реальности. Между тем, основные подходы к формализации семантики ЕЯ, популярные в 1980-е годы, не обладали выразительной силой, необходимой для эффективного описания знаний о реальности, для построения моделей концептуальной памяти и т.д.

6. Как следствие, за рубежом не разрабатывались модели соответствия “Текст – Система знаний – Семантическое представление (или представления) текста”.

7. В 1980-е годы ряд исследователей отмечали необходимость и важность моделирования процессов использования ЕЯ в общении, т.е. с учетом целей интеллектуальных систем, реализуемых в процессе общения, и их знаний о мире в целом и о другом участнике диалога (Попов 1982; Fomitchov, 1983, 1984; Narin’yani, 1984; Фомичев, 1988б; Fomichov,

1992). Однако наиболее популярные в 1980-е годы формализмы, использовавшиеся для изучения семантики ЕЯ, не предоставляли такой существенной возможности.

Представляется, что перечисленные ограничения являются наиболее важными с точки зрения проектирования семантико-синтаксических анализаторов дискурсов, относящихся к науке, технике, экономике, медицине, а также для разработки ЕЯ-интерфейсов больших баз данных и знаний.

1.4. Роль формальных систем семантических представлений с большими выразительными возможностями в проектировании лингвистических процессоров

Совокупность задач, поставленных перед теорией ЛП в начале 1980-х годов, оказалась чрезвычайно трудной. Как следствие, развитие теории ЛП в 1980-е годы сильно замедлилось. Несмотря на реализацию значительного количества проектов конструирования ЛП в разных странах мира, существенного продвижения вперед не удавалось достичь.

Главная причина этого замедления заключалась в следующем. В ЕЯ причудливым образом взаимодействуют многочисленные механизмы кодирования и декодирования информации. Поэтому часто для того, чтобы "понять" даже довольно простые для человека фразы или дискурсы, компьютер должен привлекать знания о закономерностях различных уровней языка (морфологическом, синтаксическом, семантическом), а также знания о мире и о конкретной ситуации диалога. Например, для того чтобы узнать, какие из нескольких ранее упомянутых объектов обозначаются местоимением "их", может потребоваться проведение умозаключений здравого смысла и логических рассуждений. Аналогичная ситуация имеет место и для задачи восстановления смысловой структуры фраз с пропусками слов (эллиптичных фраз) в контексте всего дискурса или всего диалога.

Поэтому, пытаясь формализовать понимание компьютером даже довольно простых текстов, исследователи быстро убеждались в том, что для решения их частных задач необходимо предварительно иметь теоретические решения, относящиеся к произвольным текстам группы естественных языков (например, русского, английского, немецкого, французского). В итоге в 1980-е годы в англоязычных публикациях даже возникла метафора "theory bottleneck" ("узкое горлышко теории"), отражающая значительные трудности создания адекватной теории понимания компьютером ЕЯ.

Наконец, несколькими группами исследователей из разных стран (в том числе и автором данной работы) была предложена идея, позволяющая найти выход из охарактеризованной тупиковой ситуации. Суть этой идеи заключается в следующем. Необходимо разработать такие формальные языки для представления знаний о мире и построения семантических представлений (СП) ЕЯ-текстов, чтобы можно было конструировать СП в виде выражений, отражающих многое структурные особенности самих текстов. Другими словами, нужны формальные языки (или формальные системы, поскольку множество их правильно построенных выражений образует язык) для описания структурированных значений (или смыслов) ЕЯ-текстов, обладающие выразительными возможностями, близкими к возможностям ЕЯ. Тогда можно будет выполнять смысловой анализ текста в два этапа:

ЕЯ-текст Т ➔ *Недоопределенное СП текста Т* ➔ *Целевое СП текста Т* .

Эту схему следует понимать следующим образом. Сначала должно быть построено промежуточное, предварительное СП текста, называемое недоопределенным семантическим представлением (НСП) рассматриваемого текста. Это выражение в большинстве случаев будет отображать смысл входного текста Т лишь частично, неполно. Например, в НСП текста Т может отсутствовать указание на конкретный объект, соответствующий конкретному вхождению в текст Т местоимения "ей" или не выбрано конкретное значение слова "станция", входящего в Т.

Однако НСП текста Т является формальным выражением, в отличие от исходного ЕЯ-текста Т. Поэтому на втором этапе обработки Т для снятия той или иной недоопределенности можно будет вызвать одну из многочисленных

специализированных процедур-"экспертов" по конкретным вопросам. Такие процедуры можно будет проектировать с примененинем формальных средств представления информации, покольку базы знаний ЛП состоят из выражений формальных языков представления знаний, а исходное НСП - вход процедуры и преобразованное НСП (в частности, совпадающее с целевым СП) являются формальными выражениями. Впервые эта идея была высказана в работах (Фомичев 1981а, 1981б; Fomitchov 1983, 1984).

С конца 1980-х годов по настоящее время идея опоры при проектировании ЛП на формальные системы семантических представлений с широкими выразительными возможностями является центральной для развития теории понимания компьютером ЕЯ. Росту популярности этой идеи способствовало появление серии публикаций по эпизодической логике (ЭЛ) (Schubert и Hwang, 1989, 2000; Hwang и Schubert, 1993а-1995), реализация на основе ЭЛ проекта TRAINS, направленного на формализацию проблемно-ориентированного диалога на естественном (английском) языке (Allen, Schubert и др. 1995), осуществление проекта машинного перевода Core Language Engine (CLE) в Кембриджском отделении (Великобритания) Стенфордского исследовательского института (Alshawi и van Eijck, 1989; Alshawi, 1990, 1992), а также осуществление в 1980-х - 1990-х годах проекта SnepS в США (Shapiro, 1996).

Отправной точкой исследований для авторов перечисленных выше работ был язык логики предикатов первого порядка. Характер предпринимавшихся усилий по расширению выразительных возможностей этого языка (точнее, класса логики предикатов первого порядка) можно проиллюстрировать следующими двумя примерами.

Пример 1. В ЕЯ-текстах встречается большое количество таких обозначений различных объектов, которые являются сочетаниями «Прилагательное + Существительное» (для русского языка) или «Артикль + Прилагательное + Существительное» (для английского, немецкого и французского языков). ЭЛ позволяет рассматривать формальные аналоги значений таких словосочетаний. Например, сочетанию «маленький дом» может соответствовать фрагмент

семантического представления (СП) текста, являющийся выражением $\exists y : [\, y\ ((атрибут, маленький) дом)]$ (Hwang и Shubert, 1993a).

В ЕЯ-текстах часто встречаются и однородные члены предложения, например, «комнату или маленький дом». ЭЛ дает возможность строить и формальные аналоги выражений такого вида. В частности, в СП текста сочетанию «комнату или маленький дом» может соответствовать формальное выражение

$\exists y : [[y \text{ комната}] \vee [y\ ((\text{ атрибут маленький}) \text{ дом})]$ (Hwang, Shubert 1993a).

В логике предикатов первого порядка нет средств построения формальных аналогов словосочетаний вида «Прилагательное + Существительное» или «Существительное1 + ‘или’ + Прилагательное + Существительное2». Поэтому различные части СП текста могут отражать различные компоненты смысла словосочетаний подобного рода.

Пример 2. В проекте «Базовый Языковой Механизм» (Core Language Engine, или CLE) формулы, используемые для построения недоопределенных семантических представлений (НСП) предложений, называются квазилогическими формами (КЛФ). В частности, выражению «the three firms» (три фирмы) будет соответствовать КЛФ (см. Alshawi 1990) q_term(<t=quant, n=plur, l=all>), S, [subset, S, q_term(<t=ref, p=def, l=the, n=number(3)>, X,[firm, X])]).

Таким образом, язык КЛФ позволяет строить формальные аналоги некоторых естественно-языковых обозначений множеств (Alshawi 1990).

Сегодня наиболее популярным за рубежом подходом к формализации семантически-ориентированных компьютерных методов анализа дискурсов является эпизодическая логика (ЭЛ). Ее создатели, Л.К. Шуберт и Ч.Х. Хуан, в указанных выше работах предложили логику семантических представлений с широкими выразительными возможностями. Создание ЭЛ представляло собою значительный вклад в формальную теорию построения и понимания ЕЯ-дискурсов и, как следствие, в формальную теорию использования ЕЯ.

В то же время анализ показывает, что выразительная сила класса формул, рассматриваемых в ЭЛ, недостаточна с точки зрения представления основных принципов использования ЕЯ интеллектуальными системами. В первую

очередь, выразительные возможности ЭЛ являются существенно ограниченными с точки зрения представления структурированных значений (СЗ) сложных целей, команд, дискурсов с описаниями множеств, дискурсов со ссылками на смысл фраз и более крупных частей текста. Кроме того, ЭЛ не предоставляет средств для рассмотрения составных обозначений понятий в качестве термов и для описания операций над понятиями. Те же ограничения (и ряд других) относятся к языкам проекта Core Language Engine и SnepS.

Логика предикатов первого порядка послужила в 1980-х годах отправной точкой для создания не только ЭЛ, но и нескольких других направлений, относящихся к области логического программирования. Одним из наиболее интересных направлений, развивающимся более 15 лет, является атрибутная логика (АЛ), или логика со значениями свойств (Johnson 1988; Carpenter 1992; Carpenter, Penn 2001). Для АЛ характерен переход от рассмотрения однородного множества логических формул, описывающих факты, ситуации, к группировке формул, характеризующих одну сущность (человека, фирму и т.д.). Перечисленные выше ограничения ЭЛ относятся и к атрибутной логике.

Проблема формализации семантики ЕЯ-текстов в течение многих лет привлекает внимание исследователей и в нашей стране. Наиболее часто это внимание было обусловлено задачей выявления и формализации таких явлений семантического уровня ЕЯ, которые можно было бы эффективно использовать для представления знаний в прикладных интеллектуальных системах. В этой связи можно отметить цикл работ Д.А. Поспелова, Ю.И. Клыкова и ряда других авторов, в которых выделяются бинарные смысловые отношения между элементами предложения для решения задач ситуационного управления (Поспелов 1975, 1981, 1986; Клыков и Горьков, 1980), Г.С. Плесневича по теории логического вывода на ассоциативных сетях и понятийно-ориентированным языкам (Плесневич 1997 - 2003), Г.С. Осипова по использованию неоднородных семантических сетей в интеллектуальных системах приобретения знаний (Осипов 1990, 1997), В.Н. Вагина по проблеме обобщения знаний, представленных семантическими сетями (Вагин 1988).

Другую часть исследователей интересовала разработка формальных средств отображения содержания (смысла) текстов, анализируемых ЛП. В работе Н.Н.

Леонтьевой (Леонтьева 1981), по-видимому, впервые в нашей стране был высказан тезис о том, что для формализации диалога, осуществляемого в ограниченных предметных областях с помощью довольно простых текстов, нужен семантический язык с выразительными возможностями, близкими к возможностям ЕЯ.

Отдельные аспекты формализации семантики ЕЯ нашли отражение в работах В.М. Брябрина, Б.Ю. Городецкого, А.К. Жолковского, А.П. Ершова, О.С. Кулагиной, Е.С. Кузина, Л.Т. Кузина, И.П. Кузнецова, Д.Г. Лахути, Н.Н. Леонтьевой, Л.И. Литвинцевой, Ю.Я. Любарского, М.Г. Мальковского, А.Г. Мацкевича, И.А. Мельчука, Л.И. Микулича, А.С. Нариньяни, Э.В. Попова, Д.А. Поспелова, В.Ш. Рубашкина, В.А. Тузова, З.М. Шаляпиной, Г.С. Цейтина, Л.Л. Цинмана и ряда других ученых.

Однако ни в какой из публикаций отечественных авторов, во-первых, не предлагается формального аппарата, удобного для отображения поверхностной смысловой структуры произвольных ЕЯ-текстов, относящихся к деловой прозе. Во-вторых, не предлагается лингвистической теории, которую можно было бы положить в основу построения математической модели, удобной для представления смысловой структуры ЕЯ-текстов деловой прозы. Таким образом, вопрос о формальных языках, удобных как для построения семантических представлений произвольных ЕЯ-текстов, относящихся к деловой прозе, так и для моделирования ЕЯ-диалога интеллектуальных систем, является чрезвычайно актуальным и остается в доступной литературе открытым (исключение составляют публикации автора данной работы).

Прогресс в решении этого вопроса означал бы существенный шаг вперед в решении фундаментальной проблемы разработки модели русского языка, сформулированной А.П. Ершовым еще в 1986 году следующим образом: “Мы хотим как можно глубже познать природу языка, и в частности русского. Одним из выражений этого познания должна стать модель русского языка. Это формальная система, которая должна быть адекватной и равнообъемной живому организму языка, но в то же время она должна быть анатомически отпрепарированной, разъятой, доступной для наблюдения, изучения и изменения” (Ершов 1986, с. 12).

Глава 2

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ДЛЯ ОПИСАНИЯ СИСТЕМЫ ПЕРВИЧНЫХ ЕДИНИЦ КОНЦЕПТУАЛЬНОГО УРОВНЯ, ИСПОЛЬЗУЕМЫХ ЛИНГВИСТИЧЕСКИМ ПРОЦЕССОРОМ

2.1. Постановка задачи

Проанализировав по доступной литературе состояние исследований в области формализации семантики ЕЯ, Перегрин (Peregrin 1990) пришел к выводу о том, что существующие логические системы не позволяют формализовать все аспекты семантики ЕЯ, являющиеся важными для проектирования ЛП. Поэтому “мы не можем использовать имеющуюся форму логики как плавильную форму, в которую любой ценой необходимо втиснуть естественный язык”, и для создания адекватной формальной теории семантики ЕЯ необходимо выполнить системное лингвистическое исследование всех компонентов ЕЯ и установить взаимосвязи между логическими подходами к формализации семантики ЕЯ и лингвистическими моделями смысла.

В сущности, к тому же выводу (но значительно раньше, на рубеже 1970-х - 1980-х годов) пришел автор данной монографии. Этот вывод явился отправной точкой для разработки излагаемой ниже постановки задачи, а также постановки задачи в главе 3.

Представляется, что границы традиционной математической логики слишком узки для того, чтобы предоставить адекватную основу для компьютерно-ориентированной формализации семантики ЕЯ. Поэтому задача создания логических основ проектирования интеллектуально мощных систем смысловой обработки ЕЯ требует не только расширения логики первого порядка, но, скорее, разработки новых математических систем, совместимых с логикой предикатов первого порядка и позволяющих формализовать логику использования ЕЯ интеллектуальными системами.

Мы будем исходить из гипотезы о том, что существует единственный ментальный уровень для представления смысла ЕЯ-выражений, который можно

назвать концептуальным уровнем, но не семантический и концептуальный уровни отдельно. Эту гипотезу подддерживают многие ученые (см., например, Meyer 1994).

Анализ показывает, что первым шагом на пути создания широко применимого и предметно-независимого математического подхода к описанию структурированных значений ЕЯ-текстов должна являться разработка формальной модели, перечисляющей первичные (т.е. не составные) единицы концептуального уровня, используемые ЛП, а также описывающей информацию, связанную с такими единицами и необходимую для соединения таких единиц в составные единицы, отображающие структурированные значения сколь угодно сложных ЕЯ-текстов.

С целью построения формальной модели, обладающей указанным свойством, был, во-первых, проведен анализ лексического состава русского, английского, немецкого и французского языков.

Во-вторых, был изучен состав первичных информационных единиц, используемых в современных языках представления знаний в прикладных интеллектуальных системах, в частности, используемых в терминологических языках представления знаний.

На основании проведенного исследования в данной главе ставится задача разработки такой предметно-независимой математической модели для описания системы первичных единиц концептуального уровня, используемых ЛП, и информации, связанной с такими единицами, которая, во-первых, конструктивно учитывает существование следующих явлений естественного языка:

На множестве понятий задана иерархия по степени их общности. Например, понятие “физический объект” является частным случаем понятия “пространственный объект”.

Нередко один и тот же предмет может быть охарактеризован с помощью нескольких понятий, ни одно из которых не является частным случаем другого; такие понятия как бы дают значения “координат объекта” по разным “семантическим осям”. Например, каждый человек является физическим объектом, способным перемещаться в пространстве. С другой стороны, каждый

человек является интеллектуальной системой, поскольку люди могут решать задачи, читать, сочинять стихи и т.д.

В русском языке есть такие слова, как “некоторый”, “определенный”, “каждый” , “какой-нибудь” , “все” , “несколько”, “большинство” и ряд других, которые в предложениях всегда присоединяются к словам и словосочетаниям, обозначающим понятия. Например, мы можем построить выражения “каждый человек” , “какой-нибудь автомобиль”, “все люди” , “несколько книг” и т.д. Аналогичные слова есть в английском, немецком, французском и многих других языках.

Во-вторых, модель должна позволять различать формальным образом обозначения первичных единиц концептуального уровня, соответствующих:

(2.1) объектам, ситуациям, процессам в реальном мире и понятиям, квалифицирующим (характеризующим) эти объекты, ситуации, процессы;

(2.2) объектам и множествам объектов;

(2.3) понятиям, квалифицирующим объекты, и понятиям, квалифицирующим множества объектов тех же видов (“корабль” и “эскадра” и т.д.);

(2.4) упорядоченным n-местным наборам различных сущностей, где n > 1 (“упорядоченная пара” и т.д.) и множествам.

В-третьих, модель должна учитывать, что совокупность первичных единиц концептуального уровня включает:

(3.1) единицы, соответствующие логическим связкам “и” , “или” , “не” и логическим кванторам существования и всеобщности;

(3.2) именам нетрадиционных функций с аргументами и/или значениями, являющимися: (3.2.1) множествами предметов, ситуаций (событий); (3.2.2) понятиями, (3.2.3) множествами понятий; (3.2.4) семантическими представлениями (СП) ЕЯ-текстов, (3.2.5) множествами СП ЕЯ-текстов;

(3.3) единицу, соответствующую слову “понятие” и отличающуюся от концептуальной единицы “понятие “; первая из упомянутых единиц вносит, например, вклад в формирование значения выражения “важное понятие, используемое в физике, химии и биологии”.

Итогом решения поставленной задачи станет определение в параграфе 2.8 класса формальных объектов, называемых концептуальными базисами. Описание класса концептуальных базисов является математической моделью, перечисляющей первичные единицы концептуального уровня, используемые ЛП, а также описывающей информацию, связанную с такими единицами и необходимую для соединения этих единиц в составные единицы, отображающие структурированные значения сколь угодно сложных ЕЯ-текстов. Данная модель является первой частью теории К-представлений (концептуальных представлений).

2.2. Базовые обозначения и вспомогательные определения

2.2.1. Общематематические обозначения

$x \in Y$ элемент x принадлежит множеству Y

$x \notin Y$ элемент x не входит в множество Y

$X \subset Y$ множество X является подмножеством множества Y

$Y \cup Z$ объединение множеств Y и Z ; $Y \cap Z$ пересечение множеств Y и Z

$Y \setminus Z$ теоретико-множественная разность множеств Y и Z, т.е. совокупность всех таких элементов x из Y , что x не входит в Z

$Z_1 \times \ldots \times Z_n$ декартово произведение множеств Z_1 , …, Z_n , где $n > 1$

$\varnothing$ пустое множество

$\forall$ для любого, для любых $\exists$ существует

$\rightarrow$ следует, влечет за собой $\Leftrightarrow$ тогда и только тогда

2.2.2. Предварительные определения и обозначения из теории формальных грамматик и языков

Определение. Алфавитом называется конечное множество символов. Если A-произвольный алфавит, то $A^+ = \{d_1,\ldots, d_n \mid n \geq 1\}$, где для $i = 1,\ldots, n$ $d_i \in A$.

Обычно вместо $d_1,\ldots,d_n$ для упрощения пишут $d_1\ldots d_n$.

Пример. $A=\{0,1\}$, 011, 11011, $0,1 \in A^+$.

Определение. Элементы множества A^+ называются непустыми цепочками (или непустыми строками) в алфавите A (над алфавитом A).

Пусть A- произвольный алфавит, d – символ из A, тогда $d^1=d$, для $n>1$ $d^n = dd\ldots d$ (n раз).

Определение. Пусть $A^* = A^+ \cup \{e\}$, где e - пустая цепочка. Тогда цепочками (или строками) в алфавите A (или над алфавитом A) называются элементы множества A^*.

Определение. Для каждого $t \in A^*$ определено значение функции *Длина (t)* (обозначаемой также через t) следующим образом: (1) $e = 0$; (2) если $t=d_1\ldots d_n$, $n \geq 1$, для $i=1,\ldots, n$ $d_i \in A$, то $t = n$.

Определение. Пусть A - произвольный алфавит. Тогда формальным языком (или, для краткости, языком) в алфавите A (или над алфавитом A) называется произвольное подмножество L множества A^*, то есть $L \in A^*$.

Пример. Пусть $A=\{0,1\}$, $L_1 = \{0\}$, $L_2=\{e\}$, $L_3=\{0^{2k}1^{2k} \mid k \geq 1\}$, тогда L_1, L_2, L_3 - языки в алфавите A.

2.2.2. Используемые определения из теории алгебраических систем

Определение. Пусть $n \geq 1$, Z – произвольное непустое множество. Тогда декартовой n-степенью множества Z называется (и обозначается через Z^n) множество Z при $n = 1$ и множество всех упорядоченных наборов вида $(x_1, x_2, \ldots, x_n)$, где $x_1, x_2, \ldots, x_n$ - элементы множества Z, при $n>1$.

Определение. Пусть $n \geq 1$, Z – произвольное непустое множество. Тогда n-арным (или n-местным) отношением на множестве Z называется произвольное подмножество R множества Z^n - декартовой n-степени множества Z. . При $n = 1$ отношение R называется *унарным отношением (*в этом случае R является произвольным подмножеством множества Z*)*, а при $n = 2$ отношение R называется *бинарным отношением* (Ершов 1970).

Пример. Пусть *Z1* – множество всех целых чисел, и *Odd* – подмножество всех нечетных чисел. Тогда *Odd* является унарным отношением на *Z1*. Пусть *Less*

является множеством всех упорядоченных пар вида *(x,y)* , где *x, y* – произвольные элементы множества *Z1*, и число *x* меньше числа *y*. Тогда *Less* – бинарное отношение на множестве *Z1*.

Очень часто вместо записи $(b,c)\in R$, где *R* – бинарное отношение на произвольном множестве *Z* , *b, c* - произвольные элементы из *Z*, используется запись *b R c* .

Определение. Пусть *Z* – произвольное непустое множество, *R* – бинарное отношение на *Z*. Тогда

(а) если для любого $a\in Z$ $(a,a)\in R$, то *R* – рефлексивное отношение;

(б) если для любого $a\in Z$ $(a,a)\notin R$, то *R* – антирефлексивное отношение;

(в) если для любых *a, b, c*$\in Z$ из $(a,b)\in R$, $(b,c)\in R$ следует, что $(a,c)\in R$, то *R* называется транзитивным отношением;

(г) если для любых *a,b* $\in Z$ из $(a,b)\in R$ следует $(b,a)\in R$, то *R* – симметричное отношение;

(д) если для любых *a, b* $\in Z$ *из* $a\neq b$ и $(a,b)\in R$ следует, что $(b,a)\notin R$, то *R* – антисимметричное отношение;

(е) если *R* - рефлексивно, транзитивно и антисимметрично, то *R* – частичный порядок на *Z* (Johnsonbaugh 2001).

Пример. Бинарное отношение *Less* из предыдущего примера является транзитивным, антирефлексивным и антисимметричным.

Пример. Пусть *Z1* – множество всех целых чисел, и *Eqless* является множеством всех упорядоченных пар вида *(x,y)* , где *x, y* – произвольные элементы множества *Z1*, и число *x* равно числу y или меньше числа *y*. Тогда *Eqless* – бинарное отношение на множестве *Z1*. Это отношение является рефлексивным, транзитивным и антисимметричным. Таким образом, отношение *Eqless* является частичным порядком на *Z1*.

Пример. Пусть *Z2*– множество всех понятий, которые обозначают транспортные средства, и *Genrel* является множеством всех упорядоченных пар вида *(x,y)* , где *x, y* – произвольные элементы множества *Z2*, и понятие *x* совпадает с понятием *y* или является обобщением понятия *y*. Например, понятие *корабль* является обобщением понятия *ледокол* ; следовательно, пара (*корабль*, *ледокол)* входит в множество *Genrel*. Очевидно, что *Genrel* – бинарное

отношение на множестве *Z2*. Это отношение является рефлексивным, транзитивным и антисимметричным. Таким образом, отношение *Genrel* является частичным порядком на *Z2*.

Определение. Пусть Z – произвольное непустое множество, R – бинарное отношение на Z. Тогда элементы $a, b \in Z$ называются сравнимыми для R, если либо $(a,b) \in R$, либо $(b,a) \in R$.

2.3. Краткая характеристика предлагаемой математической модели для описания системы единиц концептуального уровня, используемых лингвистическим процессором

С математической точки зрения, решение задачи, поставленной в параграфе 2.1, является определением нового класса формальных объектов, называемых *концептуальными базисами (к.б.)*. Отдаленным прообразом этого понятия является понятие сигнатуры алгебраической системы (Ершов, Палютин 1979).

Каждый к.б. B является упорядоченным набором вида

$$((c_1, c_2, c_3, c_4), (c_5,..., c_8), (c_9,..., c_{15}))$$

с компонентами $c_1, c_2, ..., c_{15}$, являющимися (главным образом) конечными или счетными множествами символов и выделенными элементами таких множеств. В частности, $c_1 = St$ - конечное множество символов, называемых сортами и обозначающих наиболее общие рассматриваемые понятия, $c_2 = P$ - выделенный сорт "смысл собщения", $c_5 = X$ - счетное множество цепочек, используемых как "строительные блоки" для формирования модулей знаний и семантических представлений (СП) текстов, $c_6 = V$ - счетное множество переменных, $c_8 = F$ - подмножество множества X, элементы которого называются функциональными символами.

Компонент $c_3 = Gen$ является таким бинарным отношением (частичным порядком) на St, что если пара (s,u) входит в Gen, то либо $s = u$, либо понятие, соответствующее сорту u , является конкретизацией понятия, соответствующего сорту s. Компонент $c_7 = tp$ является отображением из объединения множеств X и V в некоторое счетное множество Tps цепочек, называемых типами и характеризующих элементы из X и V.

Предположим, например, что *X* включает элементы *интс* , *дин.физ.об*, *чел*, *редсовет*, *Д.И.Менделеев*, обозначающие сорт "интеллектуальная система", сорт “динамический физический объект”, понятия “человек”, "редакционный совет" и конкретного человека – выдающегося химика Дмитрия Ивановича Менделеева. Будем рассматривать символ $\uparrow$ как индикатор почти всех типов, связанных с понятиями. Тогда значениями отображения *tp* для элементов *чел*, *редсовет* , *Д.И.Менделеев* будут элементы $\uparrow$ *интс* дин.физ.об*, $\uparrow$ {*интс* дин.физ.об* } и *интс* дин.физ.об* соответственно. Если же в качестве элемента множества *X* мы рассматриваем обозначение редакционного совета конкретного издания, то для такой информационной единицы отображение *tp* примет значение *{интс*дин.физ.об}* . Таким образом, типы помогают различать (а) объекты и понятия, характеризующие эти объекты, (б) множества и понятия, характеризующие эти множества.

Определение класса концептуальных базисов будет использовано в главе 3 следующим образом. Каждому к.б. *B* будут поставлены в соответствие три множества формул *Ls* = *Ls*(*B*), *Ts* = *Ts*(*B*), Y*s* = *Ys*(*B*) (*l*-формулы, *t*-формулы, *y*-формулы). Множество *Ls*(*B*) будет названо *стандартным К-языком в базисе B*. Его цепочки подходят для построения семантических представлений (СП) текстов на естественном языке. Каждая формула из *Ts*(*B*) имеет вид *d* & *t*, где *d* $\in$ *Ls*(*B*), *t* – тип из *Tps*(*B*). Формулы из *Ys*(*B*) имеют вид a_1 & …& a_n & *d*, где a_1, …,a_n, *d* $\in$ *Ls*(*B*), *n* имеет разные значения для разных *d*, цепочка *d* строится из a_1, …,a_n как из элементарных информационных единиц (некоторые из них могут быть немного преобразованы) однократным применением некоторого правила построения.

Например, к.б. В можно определить так, чтобы выполнялись соотношения

Ls(B) ∋ страна, нек страна , нек страна : x1,

Столица (нек страна : x1), (Столица (нек страна : x1)≡Москва);

Ts(B) ∋ страна & ↑простр.об, нек страна & простр.об,

нек. страна : x1 & простр.об, Столица (нек страна : x1) & простр. об,

(Столица (нек страна : x1) ≡ Москва) & сообщ,

Ys(B) ∋ нек & страна & нек страна, нек страна & x1 & нек страна : x1,

Столица & нек страна : x1 & Столица (нек. страна : x1),

Столица (нек страна : x1) & ≡ & Москва & (Столица (нек страна : x1) ≡ Москва) , где *сообщ=P(B)* – выделенный сорт “смысл сообщения” для рассматриваемого к.б. *B, нек* – информационная единица, соответствующая словам “некоторый” “некоторая”, “некоторое”.

2.4. Основные идеи определения класса сортовых систем

Начнем решать поставленную задачу. Будем предполагать, что необходимо построить формальное описание некоторой предметной области (ПО), и рассмотрим первые шаги в этом направлении.

Шаг 1. Введем в рассмотрение конечное множество символов, обозначающих наиболее общие понятия ПО: пространственный объект, физический объект, интеллектуальная система, натуральное число и т.д. Будем считать, что каждое такое понятие характеризует сущность, не рассматриваемую как упорядоченный набор других сущностей или как множество, состоящее из каких-то других сущностей. Обозначим это множество символов через St и будем называть его элементы сортами.

Шаг 2. Выделим в St некоторый сорт, который будем связывать с семантическими представлениями (СП) ЕЯ-текстов, выражающих отдельные высказывания либо являющихся связными повествовательн ыми текстами. Обозначим такой сорт через P и назовем его сортом «смысл сообщения». Например, для каких-то применений роль выделенного сорта P может играть цепочка сообщ. Часть формул, которые мы будем рассматривать в этой работе, представима в виде F & t , где F - СП ЕЯ-выражения, a t – цепочка, классифицирующая данное выражение. Тогда, если t = P, то подформула F интерпретируется как СП простого или сложного высказывания (другими словами, сообщения). В частности, так может интерпретироваться формула *(Вес(нек блок1 : x3) ≡ 4/тонна) & сообщ* .

Шаг 3. Введем иерархию понятий на множестве сортов *St* с помощью некоторого бинарного отношения *Gen* на *St*, т.е. выделим некоторое подмножество $Gen \subset St \times St$. Например, могут выполняться соотношения

(цел, нат), (вещ, цел), (физ. об, дин. физ. об), (простр. об, физ. об) $\in$ *Gen.*

Шаг 4. Многие объекты могут быть охарактеризованы с разных точек зрения, у них есть «координаты» по разным «семантическим осям». Например, к конкретному университету можно подъехать или подойти, поэтому каждый университет имеет семантическую координату “пространственный объект”. У университета есть руководитель (ректор) , поэтому университеты имеют семантическую координату “организация”. Наконец, университет может разработать некоторую технологию или некоторый прибор; следовательно, представляется разумным считать, что университеты имеют семантическую координату “интеллектуальная система”.

Учитывая эти соображения, введем бинарное отношение совместимости (толерантности) *Tol* на множестве *St*. Это отношение интерпретируется следующим образом: если $(s,u) \in Tol \subset St \times St$, то существует такая сущность *x* в рассматриваемой ПО, что с *x* можно связать сорт *s* по одной семантической оси и сорт *u* по другой оси, причем сорт *s* и сорт *u* не являются сравнимыми для отношения *Gen*.

Например, множества *St* и *Tol* могут быть определены так, что *Tol* включает упорядоченные пары *(простр.объект, организация), (простр.объект, интел.система), (организация, интел.система), (организация, простр.объект), (интел.система, простр.объект), (интел.система, организация)* .

Из рассмотренной интерпретации отношения *Tol* вытекают следующие свойства: (*1)* $\forall u \in St\ (u,u) \notin Tol$, т.е. *Tol* – антирефлексивное отношение; *2)* $\forall u,t \in St$ из $(u,t) \in Tol$ следует, что $(t,u) \in Tol$, т.е. *Tol* – симметричное отношение.

Сортовой системой (с.с.) будем называть произвольную четверку *S* вида *(St, P, Gen, Tol)*, компоненты которой удовлетворяют определенным условиям.

2.5. Формальное определение сортовой системы

Определение. *Сортовой системой (с.с.)* будем называть произвольную упорядоченную четверку *S* вида

$$(St, P, Gen, Tol), \qquad (2.5.1)$$

где *St* – конечное множество символов, $P \in St$, *Gen* – непустое бинарное отношение на *St*, являющееся частичным порядком на *St* (т.е. рефлексивным,

транзитивным и антисимметричным), *Tol* – бинарное отношение на *St*, являющееся антирефлексивным и симметричным, и выполняются следующие условия:

(1) *St* не включает символы *‘↑’, ‘{‘ , ‘}’ , ‘(‘ , ‘)’ , ‘,’ , [сущн], [пон], [об], [↑сущн], [↑пон], [↑об];* (2) *St \ {P}≠∅ и ∀ u∈St \ {P} u, P* – несравнимы как для отношения *Gen*, так и для отношения *Tol*; (3) *∀ t,u∈ St* из *(t,u)∈ Gen* или *(u,t)∈ Gen* следует, что *t,u* несравнимы для отношения *Tol;* (4) *∀ t1,u1∈ St, t2,u2∈St* из *(t1,u1)∈ Tol, (t2,t1)∈ Gen, (u2,u1)∈ Gen* вытекает, что *(t2,u2)∈ Tol*.

Элементы множества *St* называются сортами; *P* – сортом «смысл сообщения»; *Gen ⊂ St×St* – отношением общности; *Tol ⊂ St×St* – отношением толерантности (совместимости). Если *(u,t)∈Gen*, то будем использовать эквивалентную запись *u→t* и говорить, что *t* – конкретизация сорта *u*, а *u* – обобщение сорта *t*. Если *(s,u)∈ Tol*, то будем использовать запись *s⊥u* и говорить, что сорт *s* совместим с сортом *u*.

Символы *‘↑’, ‘{‘ , ‘}’ , ‘(‘ , ‘)’ , ‘,’ , [сущн], [пон], [об], [↑сущн], [↑пон], [↑об]* будут играть особые роли при построении из сортов цепочек, называемых типами и классифицирующих сущности, рассматриваемые в выбранной предметной области (см. параграф 2.6).

Пример. Пусть *St_0={нат, цел, вещ, простр.об, физ.об., дин.физ.об, вообр.об, интс, орг, сит, соб, мом, сообщ}.* Элементы множества *St_0* обозначают понятия и интерпретируются следующим образом: *нат* – «натуральное число», *цел* – “целое число“, *вещ* – «вещественное число», *простр.об* – «пространственный объект», *физ.об* – «физический объект», *дин.физ.об* – «динамический физический объект», *вообр.об* – «воображаемый пространственный объект» (орбиты небесных тел, геометрические фигуры), *интс* – «интеллектуальная система», *орг* – “организация“, *сит* – «ситуация», *соб* – “событие“ (т.е. динамическая ситуация), *мом* – «момент времени», *сообщ* – «семантическое представление сообщения».

Пусть *P_0=сообщ, Ge1={(u,u) | u∈St_0}* ,

Ge2={(цел, нат), (вещ, цел), (вещ, нат), (простр.об, физ.об), (простр.об, вообр.об), (физ.об, дин.физ.об), (простр.об, дин.физ.об), (сит, соб)},

$Gen_0 = Ge1 \cup Ge2$,

$T1 = \{$(интс, дин.физ.об), (интс, физ.об), (интс, простр.об), (орг, интс), (орг, физ.об), (орг, простр.об)$\}$, $T2 = \{(u,s) \mid (s,u) \in T1\}$, $Tol_0 = T1 \cup T2$.

Пусть $S_0 = (St_0, P_0, Gen_0, Tol_0)$. Тогда легко проверить, что S_0 является сортовой системой, и сорт *сообщ* является выделенным сортом «смысл сообщения» этой системы. Из определения множества Gen_0 вытекают, в частности, следующие соотношения:

вещ→цел, цел→нат, вещ→нат, простр.об.→физ.об, простр.об.→вообр.об, физ.об.→дин.физ.об; интс⊥физ.об, интс⊥дин.физ.об, интс⊥орг, физ.об⊥интс, дин.физ.об⊥интс.

2.6. Типы, порождаемые сортовыми системами, и конкретизации типов

2.6.1. Определение множества типов

Предположим, что нам необходимо описать некую предметную область (ПО), и мы решили рассматривать некоторые сущности как элементарные сущности (люди, фирмы, числа, факты, понятия и т. д.). Тогда определим составные сущности для данной области как такие сущности, которые рассматриваются как упорядоченные наборы других сущностей или как множества, состоящие из каких-то других сущностей. Будем интерпретировать понятия(другими словами, концепты) как общие описания сущностей, относящихся к некоторым различаемым людьми классам сущностей. Объекты определим как такие сущности, которые не рассматриваются как понятия. Класс объектов включает, в частности, семантические представления (СП) текстов, множества СП текстов и множества понятий.

Определим для каждой с. с. *S* множество цепочек *Tp(S)*, элементы которого назовем типами системы *S* и будем понимать их как характеристики сущностей, рассматриваемых в рассуждениях о данной области. При построении типов используются сорта из *S* и специальные символы *[сущн]*, *[пон]*, *[об]*, *[↑сущн]*, *[↑пон]*, *[↑об]*, ‘↑’, ‘{‘ , ‘}’, ‘(‘ , ‘)’, ‘,‘ (запятая). Символы *[сущн]*, *[пон]*, *[об]*

будем называть, соответственно, типом «сущность», типом «концепт» (это наиболее общая характеристика понятий) и типом «объект» (это наиболее общая характеристика сущностей, не рассматриваемых как понятия). Символ «*» будет используется для соединения нескольких совместимых сортов (т.е. сравнимых для отношения толерантности *Tol*) при построении цепочек из множества *Tp(S).* Символ '↑' будем интерпретировать как индикатор типа понятия.

Предположим, что мы используем сортовую систему S_0 , построенную в примере из параграфа 2..5. Тогда мы сможем связать с понятием "человек" тип *↑интс*дин.физ.об* из *Tp(S_0)* , с каждым конкретным человеком – тип *интс*дин.физ.об* , с понятием "студенческая учебная группа" – тип *↑{интс*дин.физ.об}* , с конкретной студенческой группой М8-05 факультета прикладной математики МИЭМ – тип *{интс*дин.физ.об}.*

Рассмотрим интерпретацию специальных символов *[сущн]*, *[пон]*, *[об]*. Формализуя рассуждения, условимся исходить из следующих рекомендаций. Если природа сущности z, рассматриваемой в рассуждении, не играет роли, то поставим в соответсвие z тип *[сущн]* в ходе рассуждения. Если же важно то, что *z* представляет собой объект, то поставим в соответсвие *z* тип *[об]*. Если, напротив, в отношении *z* важно то, что *z* является понятием, то поставим в соответсвие *z* тип *[пон]*. Назначение типов *[сущн]*, *[пон]*, *[об]* станет понятным из следующих примеров.

Пусть E_1 и E_2 соответствуют выражениям "первая сущность, упомянутая на странице 12 выпуска газеты "The Moscow Times", опубликованного 1 октября 1994 г.", и "первый объект, упомянутый на странице 12 выпуска газеты "The Moscow Times", опубликованного 1 октября 1994 г.". Тогда можно связать типы [сущн] и [об] с сущностями, на которые ссылаются в E_1 и E_2 соответственно, в случае, если мы не читали страницу 12 указанного выпуска.

Однако, прочитав эту страницу, мы узнаем, что первая сущность и первый объект, упомянутые на этой странице, — город Мадрид. Следовательно, теперь мы можем связать с упомянутой сущностью (объектом) более информативный тип простр.об («пространственный объект»).

Пусть E_3 — выражение "понятие с меткой AC060, определенное в Longman Dictionary of Scientific Usage (Moscow, Russky Yazik Publishers, 1989)". Не читая

словаря, мы можем связать с понятием, упомянутым в E_3, только тип [пон]. Но после того, как мы найдем определение с пометкой AC060, мы узнаем, что это определение понятия “трубка” (полый цилиндр с длиной много больше диаметра). Следовательно, мы можем связать с понятием, упомянутым в Е3, более информативный тип ↑физ (обозначение понятия “физический объект”).

Будем предполагать, что цепочки [↑сущн], [↑пон], [↑об] – это типы семантических единиц, соответсвующих словам “сущность”, “понятие”, “объект”. Эти типы образуют множество специальных типов Spectp.

Условимся в последующих определениях считать символами как цепочки [сущн], [пон], [об], [↑сущн], [↑пон], [↑об] , так и элементы сортовых множеств.

Определение. Пусть S — с.с. вида (1), Spectp = {[↑сущн], [↑пон]}, [↑об]}, Toptp = {[сущн], [пон]}, [об]}. Тогда через Tp(S) обозначим наименьшее множество T, удовлетворяющее следующим условиям:

(1) Spectp ∪ Toptp∪ St∪ {↑s | s∈ St} ⊆ T ; элементы множеств Spectp и Toptp называются специальными типами и верхними типами, соответственно;

(2) Если k>1, для ∀ i=1,…,k s_i∈ St, ∀ i,j=1,…,k из i≠j следует, что $s_i \perp s_j$ (т.е. цепочки s_i, s_j сравнимы для отношения совместимости Tol), то цепочка $s_1*s_2*\ldots*s_k$ и цепочка ↑ $s_1*s_2*\ldots*s_k$ входят в T;

(3) Если n > 1, для i = 1,...,n t_i∈ T \ Spectp, то цепочка вида $(t_1,...,t_n)$ входит в T;

(4)Если t∈ T \ Spectp , то цепочка {t} входит в T;

(5)Если t∈T \ (Spectp ∪ Toptp), и t начинается с символа ‘(‘ или ‘{’, то цепочка ↑t входит в T.

Множество Tp(S) называется множеством типов, порождаемых с.с. S .

Определение. Если S — с.с., то Mtp(S) = Tp(S) \ Spectp; элементы множества Mtp(S) называются основными типами (обозначение этого множества происходит от английского словосочетания main types).

2.6.2. Интерпретация определения множества типов

Сформулируем принципы установления соответствия между сущностями, рассматриваемыми в предметной области с с.с. S и типами из множества Mtp(S).

Типы понятий, в отличие от типов объектов, начинаются с символа ‘↑’. С понятием, обозначаемым сортом с, свяжем тип ↑ с. Тип {t} соответствует любому множеству сущностей типа t. Если $x_1,...,x_n$ — сущности типов $t_1,...,t_n$, тогда тип $(t_1,...,t_n)$ соответствует n-местному упорядоченному набору $(x_1,...,x_n)$. Как следствие, множества, состоящие из упорядоченных наборов с типом $(t_1,...,t_n)$, будут иметь тип $\{(t_1,...,t_n)\}$.

Пример 1. Можно связать типы из Mtp(S) с некоторыми понятиями и объектами с помощью следующей таблицы:

ПОНЯТИЕ	ТИП
понятие “множество”	↑{[сущн]}
понятие “множество объектов”	↑{[об]}
понятие “множество понятий”	↑{[пон]}
понятие “человек”	↑интс*дин.физ.об
Д.И.Менделеев	интс*дин.физ.об
понятие “студенческая группа”	↑{интс*дин.физ.об }
Группа М8-05	{интс*дин.физ.об }
понятие “пара целых чисел”	↑(цел, цел)
пара (12,144)	(цел, цел)

Можно также связать с отношением “Меньше” на целых числах тип {(*цел*, *цел*)}, с отношением “Принадлежать множеству” — тип {([сущн], {[сущн]})}, с отношением “Объект *Y* характеризуется понятием *C*” — тип {([*об*], [*пон*])}, а с отношением “Понятие *D* является обобщением понятия *C*” — тип {([*пон*], [*пон*])}.

Основная цель введения типов заключается в том, чтобы задавать семантические ограничения на атрибуты отношения, в частности, на аргументы и значения функций. Идею такого использования типов можно пояснить следующим образом. Пусть *c* – обозначение понятия (в частности, $c \in St$). Тогда через *Dt(c)* будем обозначать множество всех сущностей, которые могут быть охарактеризованы понятием *c*, и называть *Dt(c) денотатом* понятия *c*.

Например, *Dt(книга)* = множество всех книг, *Dt(человек)* = множество всех людей.

Пусть *R* – обозначение *n*-арного отношения, *n>1*. Тогда ограничение $(x_1,\ldots,x_n) \in R \Leftrightarrow x_1 \in Dt(s_1), \ldots, x_n \in Dt(s_n)$, где $s_1,\ldots, s_n \in St$, будем указывать с помощью значения некоторого отображения *tp*, определенного для *R*, следующим образом: $tp(R)=\{(s_1,\ldots, s_n)\}$. Аналогично, ограничение $(x_1,\ldots, x_n) \in R \Leftrightarrow x_1 \in Dt(c_1), \ldots, x_n \in Dt(c_n)$ будем представлять в виде $tp(R)=\{(tc_1, \ldots, tc_n)\}$, где $tc_1, \ldots, tc_n$ – типы, характеризующие сущности из рассматриваемой предметной области.

Пример 2. Семантические ограничения на атрибуты отношений *Брат*, *Расстояние* (последнее отношение является функцией, ставящей в соответствие двум пространственным объектам некоторое значение длины) можно представить следующим образом:

$$tp(Брат)=\{(интс*дин.физ.об,\ интс.*дин.физ.об)\},$$

$$tp(Расстояние)=\{(простр.об,\ простр.об,\ длин)\}.$$

2.6.3. Отношение конкретизации на множестве типов

Пусть *S* –произвольная сортовая система (с.с.). Зададим на множестве типов *Tp(S)* некоторое бинарное отношение, обозначаемое символом $\vdash$ и называемое *отношением конкретизации*. На множестве сортов *St* отношение $\vdash$ совпадает с отношением общности $\rightarrow$. Следующая система примеров демонстрирует требования к отношению $\vdash$: *[сущн]* $\vdash$ *[об]*, *[сущн]* $\vdash$ *[пон]*, *физ.об* $\vdash$ *дин.физ.об*, *дин.физ.об* $\vdash$ *интс * дин.физ.об*, *[пон]* $\vdash \uparrow$*интс*, *[пон]* $\vdash \uparrow$*интс * дин.физ.об*, *[об]* $\vdash$ *физ.об*, *[об]* $\vdash$ *{физ.об}*, *[об]* $\vdash$ *{(вещ, вещ)}*.

Основная идея определения отношения конкретизации заключается в следующем. Мы хотим, чтобы расстояние могло быть определено и между неподвижными физическими объектами, и между динамическими физическими объектами, и между воображаемыми динамическими физическими объектами. Все объекты таких видов являются частными случаями пространственных объектов. Учитывая это, будем использовать отношение конкретизации $\vdash$ следующим образом.

Пусть R – обозначение n-арного отношения, где $n>1$, и некоторое отображение tp ставит в соответствие R описание семантических ограничений на атрибуты $\{(t_1,\ldots,t_n)\}$, т.е. $tp(R)=\{(t_1,\ldots,t_n)\}$, где $n>1$, $t_1,\ldots,t_n \in Tp(S)$.

Будем полагать, что выражение $R(x_1,\ldots,x_n)$ выражает тот же смысл, что и выражение $(x_1,\ldots,x_n) \in R$. Тогда будем считать выражение $R(x_1,\ldots,x_n)$ допустимым $\Leftrightarrow$ существуют такие $u_1, \ldots,u_n \in Tp(S)$, что $\forall k=1,\ldots n \; t_k \mid\!\!- u_k$ и $x_k \in Dt(u_k)$, т.е. x_k входит в денотат понятия u_k.

Пример 3. Так как выполняются соотношения

простр. об $\rightarrow$ *вообр.простр.об, простр.об* $\rightarrow$ *физ.об, физ.об* $\rightarrow$ *дин.физ.об,*

то *простр.об* $\mid\!\!-$ *вообр.простр.об, простр.об* $\mid\!\!-$ *дин.физ.об* .

Поэтому допустимыми будут являться выражения $Pасст(x_1,x_2,l_1)$, $Pасст(z_1,z_2,l_2)$, где x_1,x_2 – обозначения двух автомобилей, z_1,z_2 – обозначения орбит двух конкретных небесных тел, и l_1, l_2 – обозначения некоторых значений длины.

Пример 4. Студенческие учебные группы является частными случаями множеств. Семантические ограничения на аргумент и значение функции “Количество элементов множества”, обозначаемой символом *Колич-элем*, можно задать соотношением *tp(Колич-элем)={({[сущн]}, нат)}* . Предположим, что база знаний интеллектуальной системы включает идентификатор студенческой группы *М8-05*, и отображение *tp* связывает с этим идентификатором тип *{интс*дин.физ.об}*. Таким образом, данная гипотетическая интеллектуальная система рассматривает объект, обозначаемый идентификатором М8-05, как некоторое множество людей (каждый человек является каг интеллектеальной системой, так и динамическим физическим объектом). Пусть *tp(14)=нат*. Так как *нат* $\rightarrow$ *нат* , то в случае выполнения соотношения *{[сущн]}* $\mid\!\!-$ *{интс*дин.физ.об}* выражение *Колич-элем(М8-05,14)* допустимо.

Определение 3. Пусть S – произвольная с.с. вида (2.5.1). Тогда *элементарными составными типами* будем называть цепочки из $Tp(S)$ вида $s_1*s_2*\ldots*s_k$, где $k>1$, для $\forall\, i=1,\ldots,k \; s_i \in St$.

Пример 5. Цепочка *интс*дин.физ.об* является элементарным составным типом для с.с. S_0 .

Определение 4. Пусть S – с.с. вида (2.5.1). Тогда через $Elt(S)$ обозначим объединение множества сортов St с множеством всех элементарных составных типов. Элементы множества Elt будем называть *элементарными типами.*

Определение 5. Если S – с.с. вида (2.5.1), $t \in Elt(S)$, то *спектр* типа t, обозначаемый через $Spr(t)$, в случае $t \in St$ является множеством $\{t\}$, а в случае $t = s_1 * s_2 * \ldots * s_k$, где $k>1$, для $\forall\ i=1,\ldots,k\ s_i \in St$, является множеством $\{ s_1,\ldots, s_k\}$.

Пример 6. Для с.с. S_0 спектр $Spr(физ.об) = \{физ.об\}$, $Spr(интс * дин.физ.об) = \{интс, дин.физ.об\}$.

Определение 6. Пусть S – с.с. вида (2.5.1), $u \in St$, t – элементарный составной тип из $Tp(S)$. Тогда тип t называется *уточнением* сорта u $\Leftrightarrow$ когда спектр $Spr(t)$ включает такой сорт w , что $u \rightarrow w$ (т.е. $(u,w) \in Gen$).

Пример 7. Пусть u = *физ.об* , t = *интс*дин.физ.об* . Тогда спектр $Spr(t)$ = { *интс, дин.физ.об* }. Поэтому из *физ.об* $\rightarrow$ *дин.физ.об* вытекает, что t - уточнение сорта u . Напомним, что в данной работе сорта считаются символами, т.е. неделимыми единицами.

Определение 7. Пусть $u \in St$, $t \in Tp(S)$, и t включает символ u. Тогда вхождение символа u называется *свободным* $\Leftrightarrow$ когда либо $t = u$, либо это вхождение u в t не является вхождением в какую-либо подцепочку вида $s_1 * s_2 * \ldots * s_k$, где $k>1$, для $\forall\ i=1,\ldots,k\ s_i \in St$, и сущеспвует такое m, $1 \leq m \leq k$, что $u = s_m$.

Пример 8. С функцией «Друзья» можно связать тип t1={(интс*дин.физ.об, {интс * дин.физ.об})}. Как первое, так и второе вхождения в цепочку t1символа дин.физ.об. не являются свободными вхождениями. С функцией “Вес множества физических объектов” можно ассоциировать тип t2={({физ.об}, (цел, кг))}; вхождения символа физ.об в t2 и в t3 = ↑физ.об (возможный тип понятия “физический объект”) являются свободными.

Определение 8. Пусть S – с.с. вида (2.5.1), тогда $Tc(S)=\{t \in Tp(S) \setminus (Spectp \cup Toptp) \mid t$ начинается с символа $\uparrow\}$, где $Spectp = \{[\uparrow сущн], [\uparrow пон]\}, [\uparrow об]\}$, $Toptp = \{[сущн], [пон]\}, [об]\}$; $Tob=Tp(S)\setminus(Spectp \cup Toptp \cup Tc(S))$.

Элементы $Tc(S)$ интерпретируются как типы понятий (кроме наиболее общего типа $[сущн]$). Элементы $Tob(S)$ интерпретируются как типы объектов (сущности, не рассматриваемые как понятия).

Определение 9. Пусть S – с.с. вида (2.5.1). Тогда преобразования $tr_1,\ldots,tr_6$, частично применимые к элементам из *Tp(S)*, задаются следующим образом:

1. Если $t \in Tp(S)$, t включает символ *[сущн]*, то tr_1 и tr_2 применимы к t. Пусть *w1* – результат замены в t произвольного вхождения символа *[сущн]* на символ *[пон]* ; *w2* – результат замены в t произвольного вхождения символа *[сущн]* на символ *[об]* . Тогда *w1* и *w2* – возможные результаты применения к цепочке t преобразований tr_1 и tr_2, соответственно.
2. Если $t \in Tp(S)$, t включает символ *[пон]*, $u \in Tc(S)$, то tr_3 применимо к t , и результат замены произвольного вхождения символа *[пон]* на u является возможным результатом применения преобразования tr_3 к t.
3. Если $t \in Tp(S)$, t включает символ *[об]*, $z \in Tob(S)$, то tr_4 применимо к t, и результат замены в t произвольного вхождения символа *[об]* на тип z является возможным результатом применения преобразования tr_4 к типу t.
4. Если $t \in Tp(S)$, t включает символ $s \in St$, $u \in St$, $(s,u) \in Gen$, то тип, получающийся из t заменой какого-либо свободного вхождения s на сорт u, является возможным результатом применения преобразования tr_5 к типу t.
5. Если $t \in Tp(S)$, $u \in St$, z – элементарный составной тип из *Tp(S)*, являющийся уточнением сорта u, w получается из t заменой произвольного свободного вхождения сорта u в цепочку t на цепочку z, то w – возможный результат применения преобразования tr_6 к цепочке t.

Пример 9. Если S_0 – построенная ранее с.с., *t1=[об]*, *t2=простр.об*, *w1=интс*дин.физ.об*, *w2= дин.физ.об* то *w1* и *w2* – возможные результаты применения преобразования tr_4 и tr_5 к *t1* и *t2*, соответственно. Если *t3={физ.об}*, *w3={интс*дин.физ.об}*, то *w3* – возможный результат применения tr_6 к *t3*.

Определение 10. Пусть S – с.с. вида (2.5.1), $t,u \in Tp(S)$. Тогда тип u называется *конкретизацией* типа t, а тип t называется *обобщением* типа u (обозначается через $t|\!-u$) $\Leftrightarrow$ либо $t=u$, либо найдутся такие $x_1,\ldots,x_n \in Tp(S)$, где $n>1$, что $x_1=t$, $x_n=u$, и для $i=1,\ldots,n-1$ найдется такое $k[I] \in \{1,\ldots,6\}$, что преобразование $tr_{k[I]}$ применимо к x_i , и x_{i+1} является возможным результатом применения преобразования $tr_{k[I]}$ к x_i.

Пример 10. Для с.с. S_0 легко проверить, что *[сущн] |–[пон]*, *[сущн] |–[об]*, *[об] |–интс*, *интс|–интс*физ.об*, *физ.об |–дин.физ.об*, *[об] |–{интс}*,

*{интс}|–{интс*дин.физ.об.}, [об] |–(вещ, вещ), [об] |–{(вещ, вещ)}, [пон] |–↑интс, [пон] |–↑интс*дин.физ.об., [пон] |–↑{интс*дин.физ.об.}*

Утверждение 2.1. Пусть *S* - произвольная сортовая система. Тогда отношение конкретизации |– на множестве типов *Tp(S)* является частичным порядком.

Доказательство. Рефлексивность и транзитивность отношения |– следуют непосредственно из определения. Антисимметричность вытекает из свойств преобразований tr_1 ,..., tr_6 . В результате применения преобразования tr_1 или tr_2 количество вхождений символа *[сущн]* уменьшается на 1. После применения преобразования tr_3 или tr_4 на 1 уменьшается количество вхождений символа *[пон]* или символа *[об]*, соответственно. Если $t1,t2 \in Tp(S)$ и тип *t2* получен из *t1* в результате однократного применения преобразования tr_5 , то это означает, что найдутся такие $s, u \in St$, что $s \neq u$, $(s,u) \in Gen$, *t1* включает символ *s*, и *t2* получается заменой некоторого вхождения символа *s* в *t1* на символ *u* . Из антисимметричности отношения *Gen* на *St* следует, что обратное преобразование *t2* в *t1* невозможно. Если тип *t2* получен из типа *t1* однократным применением преобразования tr_6 , то количество символов в *t2* больше, чем количество символов в *t1*.

2.7. Концептуально-объектные системы

Предположим, что для описания какой-то предметной области (ПО) мы выбрали некоторую с.с. *S* вида (2.5.1). Тогда на следующем шаге выберем некоторое множество *X*, состоящее из таких элементарных информационных единиц, с помощью которых мы будем описывать сообщения, команды и вопросы, относящиеся к рассматриваемой ПО; это множество *X* будем называть *первичным информационным универсумом*. Затем выберем *V* – некоторое счетное множество символов, называемых *переменными* и используемых в качестве меток разнообразных сущностей, в том числе в качестве меток СП текстов и фрагментов СП текстов.

Далее зададим отображение $tp: X \cup V \to Tp(S)$ из объединения $X \cup V$ в множество типов, порождаемых с.с. S, тогда каждая переменная и каждая сущность получат тип. На последнем шаге выделим некоторое подмножество *F* множества

X так, что элементы F будут являться обозначениями функций, рассматриваемых в данной ПО. Тогда набор (X,V, tp,F) будет являться концептуально-объектной системой, согласованной с с.с. S.

Определение 1. Пусть S – с.с. вида (2.5.1). Тогда произвольную упорядоченную четверку Ct вида

$$(X, V, tp, F) \qquad (2.7.1)$$

назовем *концептуально-объектной системой (к.о.с.),* согласованной с с.с. S (или к.о.с. для S) ⇔ когда выполняются следующие условия:

(1) X, V – счетные непересекающиеся множества символов; tp - отображение вида $X \cup V \to Tp(S)$;

(2) (2) $F \subset X$, для каждого $r \in F$ цепочка $tp(r)$ начинается с подцепочки ‘{(‘ и заканчивается подцепочкой ‘)}’;

(3) (3) $St \subset X$, и для любого $s \in St$ $tp(s) = \uparrow s$;

(4) (4) $\{ v \in V \mid tp(v) = [сущн] \}$ – счетно.

Множество X называется *первичным информационным универсумом*, элементы множеств V и F называются, соответственно, *переменными* и *функциональными символами*. Если элемент $d \in X \cup V$, $tp(d)=t$, то будем говорить, что t – тип элемента d.

Пример. Построим некоторую к.о.с. Ct_0 для с.с. S_0. Пусть N – множество всех цепочек из цифр ‘0’, ‘1’, … , ‘9’, таких, что если первый символ цепочки 0, то и вся цепочка – 0. Будем полагать, что символы

чел, химик, биолог, студ.гр, тур.гр, П.Сомов, А.Зубов, И.Семенов, Друзья, Колич, Меньше, Знает, Явл1, Сейчас, Раньше, Включить1, Элем

являются соответственно обозначениями понятий “человек“, “химик”, “биолог”, “студенческая группа”, “туристическая группа”, трех конкретных людей, функции “Друзья”, функции “Количество элементов множества”, отношения “Меньше” на множестве вещественных чисел, отношения “В памяти некоторой интеллектуальной системы X1 в момент времени X2 имеется концептуальное представление некоторого сообщения X3”, отношения “Некоторый объект X1 характеризуется понятием X2” (пример реализации в тексте: “П.Сомов является химиком”), текущего момента времени, отношения “Раньше” на множестве моментов времени, отношения “Некоторая интеллектуальная система X1

включает сущность X2 в момент X3 в состав множества сущностей X4", отношения "Элемент множества", отношения "Подмножество". Символ *понятие* будем интерпретировать как информационную единицу, соответствующую словам "понятие" и "концепт".

Пусть *U1={чел, химик, биолог, студ.гр, тур.гр, П.Сомов, А.Зубов, И.Семенов, Друзья, Колич, Меньше, Знает, Явл1, Сейчас, Раньше, Включить1, Элем, понятие}*.

Зададим отображение *t1* из *U1* в $Tp(S_0)$ следующей таблицей:

x	*t1(x)*
чел, химик, биолог	*↑интс*дин.физ.об*
студ.гр, тур.гр	*↑{интс*дин.физ.об}*
П.Сомов, А.Зубов, И.Семенов	*интс*дин.физ.об*
Друзья	*{ (интс*дин.физ.об, {интс*дин.физ.об}) }*
Колич	*{ ({[сущн]}, нат) }*
Меньше	*{ (вещ, вещ) }*
Знает	*{ (интс, мом, сообщ) }*
Явл1	*{ ([об], [пон]) }*
Сейчас	*мом*
Раньше	*{ (мом, мом) }*
Включить1	*{ (интс, [сущн], мом, {[сущн]}) }*
Элем	*{ ([сущн], {[сущн]}) }*
понятие	*[↑пон]*

Табл. 2.1. Примеры соответствий между сущностями и типами

Будем полагать, что *АО_"Салют"*, *АО_"Старт"*, *НПО_"Радуга"* являются обозначениями организаций, *Поставщики*, *Персонал*, *Директор* – обозначения функций "Множество всех поставщиков данной организации", "Множество всех сотрудников данной организации" и "Директор данной организации",

соответственно. Пусть $U2$ = { *АО_"Салют", АО_"Старт", НПО_"Радуга", Поставщики, Персонал, Директор*}, и отображение $t2$ из $U2$ в $Tp(S_0)$ задается следующими условиями:

$t2$(*АО_"Салют"*) = $t2$(*АО_"Старт"*)= $t2$(*НПО_"Радуга"*)

= *орг *простр.об*интс*,

$t2$(*Поставщики*) = {(*орг*, {*орг*})}, $t2$(*Персонал*) = {(*орг*, {*интс*дин.физ.об*})},

$t2$(*Директор*) = {(*орг, интс*дин.физ.об*)} .

Пусть Vx={*x1, x2, …*}, Ve = {*e1, e2, …*} , Vp={*P1, P2, …*} ,

$Vset$ = {*S1, S2, …*} , $V_0 = Vx \cup Ve \cup Vp \cup Vset$, $X_0 = St_0 \cup N \cup U1 \cup U2$,

и отображение $tp_0 : X_0 \cup V_0 \to Tp(S_0)$ задается следующими соотношениями:

$d \in St_0$ => $tp_0(d) = \uparrow d$; $d \in Nat$ => $tp_0(d)$=*нат*; $d \in U1$ => $tp_0(d) = t1(d)$;

$d \in U2$ => $tp_0(d) = t2(d)$; $d \in Vx$ => $tp_0(d)$=[*сущн*]; $d \in Ve$ => $tp_0(d)$=*сит*;

$d \in Vp$ => $tp_0(d)$=*сообщ*, $d \in Vset$ => $tp_0(d)$= {[*сущн*]}.

Пусть F_0={*Друзья, Колич, Поставщики, Персонал, Директор*},

$Ct_0 = (X_0, V_0, tp_0, F_0)$, тогда нетрудно проверить, что Ct_0 – к.о.с. для с.с. S_0.

2.8. Системы кванторов и логических связок. Концептуальные базисы

Предположим, что мы определили с.с. S вида (2.5.1) и к. о. с. Ct вида (2.7.1) для описания рассматриваемой предметной области. Тогда предлагается выделить в первичном информационном универсуме X два непересекающихся и конечных (следовательно, непустых) подмножества Int_1 и Int_2 следующим образом: выделим в St два сорта int_1 и int_2 и предположим, что для $m=1,2$ $Int_m = \{x \in X | tp(x) = int_m\}$. Элементы Int_1 соответствуют значениям выражений "каждый", "какой-то", "некоторый", "произвольный" и т. д. в случаях, когда эти выражения являются частями групп слов, и эти группы связаны с единственным числом. Элементы Int_2 интерпретируются как семантические единицы, соответствующие выражениям "все", "несколько", "почти все", "многие" и т. д. ; минимальное требование к Int_2 заключается в том, чтобы Int_2 содержало семантическую единицу, соответствующую слову "все". Пусть Int_1 содержит выделенный элемент *ref*, рассматриваемый как аналог слова "некоторый" в

смысле “какой-то вполне определенный” (но, возможно, неизвестный). Если *Ct* — к. о. с. вида (2), $d \in X$, *d* обозначает понятие, и семантическое представление (СП) текста включает подцепочку вида *ref d* (например, цепочку *нек человек*, где *ref = нек, d = человек*), тогда будем полагать, что эта подцепочка обозначает некоторую конкретную сущность (но не произвольную), которая характеризуется понятием *d*.

Кроме того, будем предполагать, что X содержит элементы ‘≡’, ‘¬’, ‘∧’, ‘∨’, понимаемые как связки “тождественно”, “не”, “и”, “или”, и элементы ∀ и ∃, понимаемые как квантор всеобщности и квантор существования. Наконец, будем считать, что множество *St* включает выделенные сорта *eqv, neg, binlog, ext* , интерпретируемые, соответственно, как типы (а) связки ‘≡’ , (б) связки отрицания ‘¬ё’ (в) связок ‘∧’, ‘∨’ (конъюнкция и дизъюнкция) , (г) квантора всеобщности и квантора существования.

Эти предположения в наглядной форме отражает рисунок 2.1. На этом рисунке *[сущн]* – это тип “сущность”; элементы *интс, дин.физ.об, нат, сит, сообщ* – сорта “интеллектуальная система”, “динамический физический объект”, “натуральное число”, “ситуация”, “смысл сообщения”; *чел, нек, произвол, определ, все, нескол* – информационные единицы, соответствующие словам “человек”, “некоторый” (“некоторая”, “некоторое”), “произвольный” (“любой”), “определенный”, “все”, “несколько”; *Колич* – обозначение функции “Количество элементов множества”. Элемент *нек* интерпретируется как квантор референтности.

Определение 1. Пусть *S* — с.с. вида (2.5.1), *Ct* — к. о. с. вида (2.7.1) для *S*, $ref \in X$, различные элементы $int_1, int_2, eqv, neg, binlog, ext$ — некоторые выделенные сорта из $St \setminus \{P\}$, и каждая пара их несравнима для отношения общности *Gen* и несравнима для отношения совместимости *Tol.* Тогда упорядоченная семерка *Ql* вида

$$(int_1, int_2, ref, eqv, neg, binlog, ext) \qquad (2.8.1.)$$

называется *системой кванторов и логических связок (с. к. л. с.) для S и Ct* ⇔когда выполнены следующие условия:

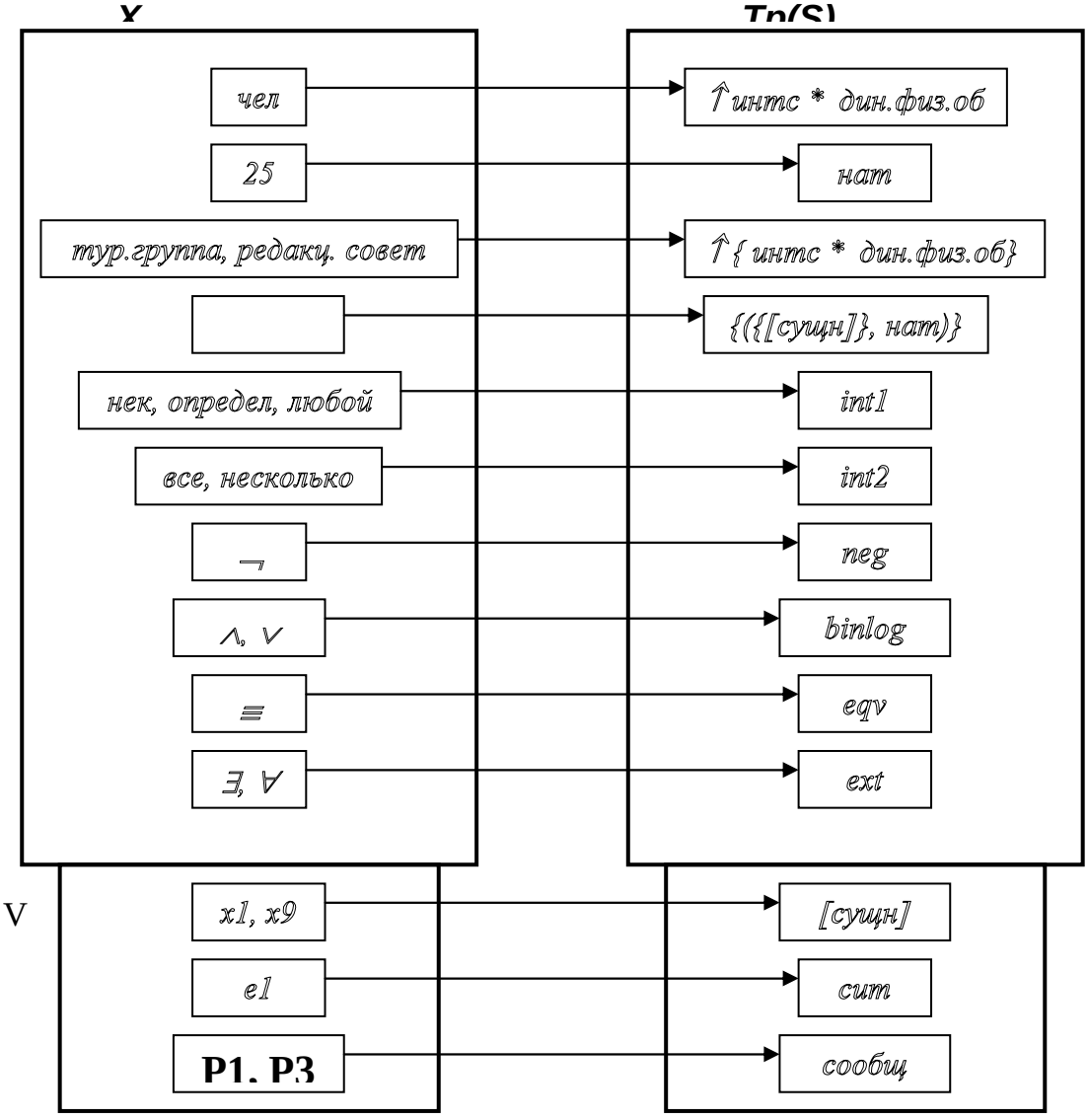

Рис. 2.1. Иллюстрация соответствия между элементами первичного информационного универсума *X*, переменными из множества *V* и их типами из множества *Tp(S)*, где *S* – сортовая система.

1. Для каждого $m = 1,2$ множество $Int_m = \{x \in X | tp(x) = int_m\}$ конечно; $ref \in Int_1$; Int_1 и Int_2 не пересекаются.
2. $\{ \equiv, \neg, \wedge, \vee, \forall, \exists \} \subset X$; кроме того, $tp(\equiv) = eqv$, $tp(\neg) = neg$, $tp(\wedge) = tp(\vee) = binlog$, $tp(\forall) = tp(\exists) = ext$.
3. Не найдется такого $d \in X \setminus (Int_1 \cup Int_2 \cup \{ \equiv, \neg, \wedge, \vee, \forall, \exists \})$ и такого $s \in \{ int_1, int_2, eqv, neg, binlog, ext\}$, что $tp(d)$ и s сравнимы для отношения общности *Gen* или сравнимы для для отношения совместимости *Tol*.
4. Для каждого $u \in \{ int_1, int_2, eqv, neg, binlog, ext \}$ сорт u и сорт P несравнимы для отношения общности Gen и несравнимы для отношения совместимости *Tol*.

Элементы Int_1 и Int_2 называются *интенсиональными кванторами*, *ref* называется *квантором референтности*, $\forall$ и $\exists$ называются *экстенсиональными кванторами*.

Пример 1. Пусть $S_0=(St_0, сообщ, Gen_0, Tol_0)$ – с.с., построенная в параграфе 2.5, $Ct_0=(X_0, V_0, tp_0, F_0)$ – к.о.с. для S_0, построенная в параграфе 2.7.. *Пусть* $Str=\{кв.инт1, кв.инт2, экв, не, бин, экст\}$, $Gen_1=Gen_0 \cup \{(s,s) / s \in Str\}$, $St_1=St_0 \cup Str$, $S_1=(St_1, сообщ, Gen_1, Tol_0)$. Тогда, очевидно, S_1 - сортовая система.

Определим теперь некоторую концептуально-объектную систему Ct_1 и некоторую систему кванторов и логических связок Ql_1. Пусть $Z=\{нек, все, \equiv, \neg, \wedge, \vee, \forall, \exists\}$, $X_1=X_0 \cup Str \cup Z$. Зададим отображение tp_1 из $X_1 \cup V_0$ в $Tp(S_1)$ следующим образом:

$u \in Str \Rightarrow tp_1(u) = \uparrow u$; $d \in X_0 \Rightarrow tp_1(u)=tp_0(u)$;

$tp1(нек)=кв.инт1$, $tp1(все)=кв.инт2$, $tp1(\equiv)=экв$, $tp1(\neg)=не$,

$tp1(\wedge) = tp1(\vee) = бин$, $tp1(\forall)=tp1(\exists)=экст$.

Пусть $Ct_1=(X_1, V_0, tp_1, F_0)$, $Ql_1= (кв.инт1, кв.инт2, нек, экв, не, бин, экст)$. Тогда легко проверить, что Ct_1 – к.о.с. для S_1, Ql_1 – с.к.л.с. для $S1$. В системе Ql_1 информационная единица *нек* интерпретируется как квантор референтности *ref* (т.е. как обозначение значения слов “некоторый”, “некоторая”, ‘некоторое”).

Определение 2. Упорядоченная тройка B вида

$(\ S, Ct, Ql\ )$ (2.8.2)

называется *концептуальным базисом (к.б.)* $\Leftrightarrow S$ – с.с., Ct – к.о.с. вида (2.7.1) для S, Ql – система кванторов и логических связок (с.к.л.с.) для S и Ct, и $(X \cup N) \cap \{$‘,’, ‘(‘, ‘)’, ‘:’, ‘*’, ‘<’, ‘>’, ‘&’$\} = \varnothing$.

Обозначим через *S(B), Ct(B), Ql(B)* компоненты произвольного к.б. вида (2.8.2). Каждый компонент *h* систем видов (2.5.1), (2.7.1), (2.8.1) будет обозначаться через *h(B)*. Например, сорт «смысл сообщения» будет обозначаться через *P(B)*, первичный информационный универсум *X* через *X(B)*, множество переменных *V* через *V(B)*, множество функциональных символов *F* через *F(B)*.

Каждый концептуальный базис будет интерпретироваться как формальное перечисление:

(а) первичных единиц, необходимых для построения семантических представлений (СП) ЕЯ-текстов, для описания знаний о реальности и для представления целей интеллектуальных систем; (б) информации, связанной с этими единицами и необходимой для построения СП текстов, формирования фрагментов знаний и представления целей интеллектуальных систем.

Пример 2. Пусть St_1, Ct_1, Ql_1 – соответственно с.с., к.о.с., с.к.л.с., определенные выше. Тогда очевидно, что упорядоченная тройка $B_1 = (\ St_1, Ct_1, Ql_1\ )$ является концептуальным базисом., и $St(B_1)=St_1$, $P(B_1)=$ *сообщ*, $X(B_1)=X_1$, $V(B_1)=V_0$.

Из этого примера следует, что множество всех концептуальных базисов не является пустым, поскольку мы построили формальный объект B_1, являющийся концептуальным базисом.

Введенные понятия дадут возможность в главе 3 для каждого к.б. *B* задать множество формул *Ls(B)*, удобных для описания содержания (т.е. структурированных значений) ЕЯ-текстов, представления знаний о мире и целей интеллектуальных систем.. Множество формул *Ls(B)* будет названо стандартным К-языком (концептуальным языком), порождаемым базисом *B*.

2.9. Обсуждение разработанной математической модели для описания системы первичных единиц концептуального уровня, используемых лингвистическим процессором

2.9.1. Особенности модели с математической точки зрения

По своей форме разработанная математическая модель для описания системы первичных единиц концептуального уровня, используемых лингвистическим процессором (ЛП), является оригинальной. Рассмотрим отличительные черты построенной модели, представляющиеся наиболее важными как с математической точки зрения, так и с точки зрения использования модели при проектировании ЛП.

1. Конструктивно учитывается существование иерархии понятий: для этого на множестве сортов St задается частичный порядок Gen, называемый отношением общности.
2. Многие сущности, рассматриваемые в той или иной предметной области, могут быть охарактеризованы с разных точек зрения. Например, люди являются, с одной стороны, интеллектуальными системами (поскольку могут читать, решать задачи и т.д.), но, с другой стороны, являются физическими объектами, способными перемещаться в пространстве. Поэтому многие понятия как бы имеют “координаты” по разным “семантическим осям”. Для учета этого важного явления в модель вводится бинарное отношение Tol (отношение совместимости, или толерантности) на множестве сортов. Накопленный опыт показал, что эта оригинальная черта модели является чрезвычайно важной для разработки алгоритмов семантико-синтаксического анализа текстов: дело в том, что появление одного и того же слова в несхожих контекстах может объясняться реализацией в этих контекстах различных “семантических координат” данного слова.
3. Фраза “это понятие используется в физике и химии” (относящаяся, например, к понятию “молекула”) для человека, владеющего русским языком, является очень простой. Между тем, смысловая структура данной

фразы не может быть адекватно отображена средствами основных известных подходов к формализации семантики ЕЯ. Причина заключается в том, что такие подходы не предлагают формального аналога информационной единицы, соответствующей слову “понятие”. Модель, построенная выше, во-первых, рассматривает специальный базовый тип [^пон], интерпретируемый как тип информационной единицы, соответствующей слову “понятие”. Во-вторых, компонент концептуально-объектной системы *Ct* вида (1.2), обозначаемый через X и называемый первичным информационным универсумом, может включать символ, интерпретируемый как информационная единица, соответствующая слову “понятие (см. пример в параграфе 1.7). Данная черта модели важна для разработки ЛП, обрабатывающих научные и научно-технические тексты, а также ЛП прикладных интеллектуальных систем, извлекающих знания из энциклопедических словарей или пополняющих электронные энциклопедических словари.

4. Одной из наиболее важных отличительных черт построенной модели является оригинальное определение множества типов, порождаемого произвольной сортовой системой, где типы рассматриваются как формальные характеристики сущностей, относящихся к выбранной предметной области. В соответствии с этим определением, (а) типы объектов из предметной области по своей форме отличаются от типов понятий, квалифицирующих данные объекты, (б) типы объектов по своей форме отличаются от типов множеств, состоящих из таких объектов, (в) типы понятий, обозначающих объекты, по своей форме отличаются от типов понятий, квалифицирующих множества данных объектов (например, тип понятия “человек” отличается от типа понятий “ученый совет”, “студенческая группа”) и т.д.
5. Модель связывает типы и с именами функций. При этом определение множества типов позволяет разумным образом связать типы с целым рядом довольно нестандартных, но практически важных функций. В частности, к ним относятся функции, значениями которых являются: (а) множество понятий, поясняемых в энциклопедическом словаре, (б)

множество понятий, входящих в определение данного понятия в данном словаре, (в) множество известных определений данного понятия, (г) количество элементов данного множества, (д) множество поставщиков данного предприятия, (е) множество сотрудников данной организации.

2.9.2. Сравнение модели с другими подходами к описанию первичных единиц концептуального уровня

Сравним построенную модель с подходами к описанию первичных единиц концептуального уровня, предлагаемыми логикой предикатов первого порядка, теорией представления дискурсов, теорией обобщенных кванторов, теорией концептуальных графов и эпизодической логикой.

В стандартной логике предикатов рассматриваются неструктурированные множества констант, функциональных символов и предикатных символов. В многосортных логиках предикатов множество констант разбито на непересекающиеся классы, каждый из которых характеризуется некоторым сортом. Разработанная выше модель предоставляет, в частности, следующие дополнительные возможности по сравнению с многосортными логиками предикатов: (1) благодаря введению отношения совместимости в качестве компонента сортовой системы, с первичной единицей концептуального уровня можно связать не только один, но и, во многих случаях, несколько сортов, как бы являющихся “координатами по ортогональным семантическим осям” сущностей, квалифицируемых или обозначаемых такими единицами; (2) “привязывание” типов к первичным информационным единицам означает, что множество таких единиц обладает развитой структурой; в частности, типы позволяют формально различать (а) типы объектов из предметной области и типы понятий, квалифицирующих данные объекты, (б) типы объектов и типы множеств, состоящих из таких объектов, (в) типы понятий, обозначающих объекты, и типы понятий, квалифицирующих множества данных объектов, (3) рассмотрение единиц концептуального уровня, соответствующих словам “некоторый”, “определенный”, “какой-нибудь”, “все”, “большинство”, “несколько”.

Кроме того, рамки построенной модели позволяют рассматривать функции, аргументами и.или значениями которых могут быть семантические представления (СП) высказываний и повествовательных текстов. Например, такая функция может ставить в соответствие каждому понятию, определяемому в энциклопедическом словаре, формулу – СП определения данного понятия. Между тем, в логике предикатов первого порядка аргументами и значениями функций могут быть только термы, но не формулы. Аналогичные ограничения должны выполняться и для атрибутов отношений, т.е. для аргументов предикатов.

Теорию представления дискурсов (ТПД) можно рассматривать как один из вариантов логики предикатов первого порядка, сочетающий использование формул и двумерных диаграмм для более наглядного представления информации. Поэтому перечисленные преимущества разработанной модели для описания системы первичных единиц концептуального уровня, используемых ЛП, относятся и к ТПД.

В теории обобщенных кванторов (ТОК), как и в построенной модели, рассматриваются единицы концептуального уровня, соответствующие словам “некоторый”, “определенный”, “все”, “большинство”, “несколько”. Однако все остальные перечисленные преимущества модели по сравнению с логикой предикатов первого порядка являются одновременно и преимуществами по сравнению с подходом ТОК.

В отличие от логики предикатов первого порядка, нотация теории концептуальных графов (ТКГ) позволяет различать обозначения конкретных объектов (конкретных компьютеров, предприятий, городов и т.д.) и обозначения понятий, квалифицирующих эти объекты (“компьютер”, “предприятие”, ‘город’). Остальные же перечисленные выше свойства модели являются преимуществами и по сравнению с ТКГ.

Наконец, указанные преимущества по сравнению с логикой предикатов первого порядка являются одновременно и преимуществами по сравнению с подходом к структурированию совокупности первичных единиц концептуального уровня, предлагаемым эпизодической логикой.

Очевидно, что если какая-либо модель предназначена как для описания системы первичных единиц концептуального уровня, используемых ЛП, так и для представления информации, связанной с этими единицами и определяющей возможности их соединения в правильные составные структуры, то подобная модель предлагает некоторый способ концептуальной структуризации рассматриваемых предметных областей. Поэтому на основании проведенного анализа можно придти к заключению, что построенная выше модель предлагает более “тонкоячеистую” структуризацию предметных областей по сравнению с основными известными подходами к формализации семантики ЕЯ, значительно увеличивает “разрешающую способность” формального инструментария, предназначенного для исследования различных предметных областей.

В конце 1990-х – начале 2000-х годов большую актуальность приобрели исследования по разработке онтологий различных предметных областей, т.е. по созданию формальных описаний систем понятий, относящихся к выбранной области, вместе с их определениями и фрагментами знаний, “привязанных” к понятиям. Первым шагом в каждом проекте такого рода является выбор некоторой начальной (другими словами, базовой) структуризации рассматриваемой предметной области или группы областей.

Представляется, что построенная в данной главе математическая модель для описания системы первичных единиц концептуального уровня, используемых ЛП, и для представления информации, связанной с этими единицами, может найти применение в проектах разработки более совершенных онтологий в произвольных предметных областях, поскольку разработанная модель предлагает формальный инструментарий с наибольшей “разрешающей способностью” по сравнению с другими известными подходами к формализации семантики ЕЯ и, как следствие, к концептуальной структуризации предметных областей.

Глава 3

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ДЛЯ ОПИСАНИЯ СТРУКТУРИРОВАННЫХ ЗНАЧЕНИЙ ПРЕДЛОЖЕНИЙ И СВЯЗНЫХ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

3.1. Постановка задачи

В параграфе 2.1. была обоснована актуальность разработки широко применимых формальных языков для построения семантических представлений (СП) ЕЯ-текстов, выразительные возможности которых позволяют отображать многие особенности поверхностной структуры предложений и связных ЕЯ-текстов. В качестве первого шага на пути разработки определения такого класса формальных языков в главе 2 был определен класс формальных объектов, называемых концептуальными базисами.

В данной главе ставится задача разработки математической модели для описания структурированных значений предложений и связных естественно-языковых текстов.

По своей форме модель должна являться описанием такого соответствия между произвольным концептуальным базисом *B* и некоторым множеством формул *Forms(B)*, чтобы класс формальных языков {*Forms(B)*, где *B* – к.б.} был удобен для построения СП фраз и связных ЕЯ-текстов, отражающих многие особенности поверхностной структуры текстов.

С целью выработки критериев для построения такой модели был проведен системный анализ структурных особенностей (а) текстов на русском, английском, немецком и французском языках, (б) ряда искусственных языков, используемых для построения семантических представлений текстов лингвистическими процессорами, (в) выражений искусственных языков представления знаний в прикладных интеллектуальных системах (в частности, терминологических языков представления знаний).

Проведенный анализ показал, что есть несколько важных аспектов формализации семантики ЕЯ, которые до недавнего времени недооценивались или игнорировались большей частью исследователей. В частности, это относится к формальному исследованию смысловых структур (а) повествовательных текстов, включающих описания множеств; б) дискурсов со ссылками на смысл предложений и более крупных частей текста; (в) фраз, где логические связки “и”, “или” используются нетрадиционной способами и соединяют не фрагменты, выражающие высказывания, а описания объектов, множеств, понятий; (г) фраз с придаточными определительными и причастными оборотами; (д) фраз со словами "понятие", "термин".

Кроме того, несколько наиболее популярных подходов к математическому изучению семантики ЕЯ не принимают или недостаточно принимают во внимание роль знаний о мире в понимании ЕЯ и, следовательно, не изучают проблем формального описания фрагментов знаний (определения понятий и т. д.). Например, это относится к весьма популярной теории представления дискурсов (Kamp 1981; Kamp, Reyle 1993, 1996; van Eijck, Kamp 1996).

Надо добавить, что тексты имеют авторов, могут быть опубликованы тем или другим источниками, могут вводиться с того или другого терминалов и т. д. Информация об этих внешних связях текстов может быть важна для их смысловой интерпретации. Поэтому целесообразно рассматривать текст как некий структурированный объект, обладающий поверхностной структурой T, множеством значений S (в большинстве случаев S состоит из одного значения), соответствующим T, и некоторыми значениями V_1 ,..., V_N, обозначающими автора (авторов) T, дату написания (или коррекции) T, указывающими новую информацию в T и т. д. Но наиболее популярные подходы к математическому исследованию ЕЯ не предусматривают формальных средств для представления текстов как структурированных объектов подобного рода.

На основании проведенного системного исследования поставим задачу построения такой модели, чтобы ее формальные средства позволяли нам следующее:

(Свойство 1): Строить обозначения структурированных значений (СЗ) как фраз, выражающих высказывания, так и повествовательных текстов; такие

обозначения обычно называют семантическими представлениями (СП) ЕЯ-выражений.

(Свойство 2): Строить и различать формальными средствами обозначения СЗ повествовательных текстов, СЗ целей (выраженных неопределенными формами глаголов с зависимыми словами, таких как "окончить с отличием МГУ, подготовить и защитить кандидатскую диссертацию по биохимии") и СЗ вопросов.

(Свойство 3): Строить и различать обозначения единиц, соответствующих (а) объектам, ситуациям, процессам в реальном мире и (б) понятиям, квалифицирующим (характеризующим) эти объекты, ситуации, процессы.

(Свойство 4): Строить и различать обозначения: (3.1) объектов и множеств объектов; (3.2) понятий и множеств понятий; (3.3) СП текстов и множеств СП текстов.

(Свойство 5): Различать формальным образом понятия, квалифицирующие объекты, и понятия, квалифицирующие множества объектов тех же видов.

(Свойство 6): Строить составные обозначения понятий, т. е. строить формулы, отражающие поверхностно-семантическаю структуру ЕЯ-выражений, подобных выражению “человек, окончивший МГУ имени М.В. Ломоносова и являющийся биологом или химиком”.

(Свойство 7): Строить объяснения более общих понятий с помощью менее общих; в частности, строить цепочки вида (a=Des(b)), где a обозначает некоторое понятие, которое необходимо объяснить, a Des(b) обозначает описание некоторой конкретизации известного понятия b.

(Свойство 8): Строить обозначения упорядоченных n-местных наборов различных сущностей, где $n > 1$.

(Свойство 9): Строить (9.1) формальные аналоги составных обозначений множеств ("эта группа, состоящая из 12 туристов, являющихся химиками или биологами" и т.п.), (9.2) обозначения множеств упорядоченных наборов сущностей, (9.3) обозначения множеств, состоящих из множеств, и т.д.

(Свойство 10): Описывать теоретико-множественные отношения и операции над множествами.

(Свойство 11): Строить обозначения СЗ фраз, содержащих, в частности:

(11.1) слова “произвольный”, “некоторый”, “все”, “каждый”, и т. д.;

(11.2) выражения, полученные применением связок “и”, “или” к обозначениям (11.2а) предметов, событий; (11.2б) понятий; (11.2в) множеств;

(11.3) выражения , где связка “не” стоит непосредственно перед обозначением предмета, события и т. д.; (11.4) косвенную речь;(11.5) причастные обороты и придаточные определительные предложения;

(11.6) слова "понятие", "термин".

(Свойство 12): Строить обозначения СЗ дискурсов со ссылками на упомянутые объекты.

(Свойство 13): Указывать явно в СП дискурсов причинно-следственные и временные отношения между описываемыми ситуациями (событиями).

(Свойство 14): Описывать СЗ дискурсов со ссылками на смысл фраз и более крупных фрагментов рассматриваемых текстов.

(Свойство 15): Выражать суждения о тождественности двух сущностей.

(Свойство 16): Строить формальные аналоги формул логики предикатов первого порядка с кванторами существования и/или всеобщности.

(Свойство 17): Рассматривать нетрадиционные функции (и другие нетрадиционные отношения) с аргументами и/или значениями, являющимися: (17.1) множествами предметов, ситуаций (событий); (17.2) множествами понятий; (17.3) множествами СП текстов.

(Свойство 18): Строить концептуальные представления текстов как информационные объекты, отражающие не только смысл, но и значения внешних характеристик текста: авторов, дату, области применения результатов и т. д.

Эта постановка задачи отражена в публикациях (Фомичев 1981а, 1981б, 1983, 1988; 2002б, в; Fomitchov 1984; Fomichov 1992, 1996a, б, 2002b).

3.2. Краткая характеристика предлагаемого решения поставленной задачи

3.2.1. Краткая характеристика новых правил построения формул

В данной главе произвольному концептуальному базису (к.б.) *B* будут поставлены в соответствие три множества формул *Ls = Ls*(*B*), *Ts = Ts*(*B*), Y*s = Ys*(*B*) (*l*-формулы, *t*-формулы, *y*-формулы). Объединение этих множеств будет обозначено через *Forms(B).* Множество *Ls*(*B*) будет названо *стандартным К-языком в к.б. B.* Концептуальный базис *B* оказывается возможным определить таким образом, что цепочки языка *Ls = Ls*(*B*) будет удобно использовать для описания структурированных значений (другими словами, смысловых структур) ЕЯ-текстов, представления знаний о мире и представления целей интеллектуальных систем.. Другими словами, цепочки из языка *Ls = Ls*(*B*) окажется удобным использовать для построения семантических представлений (СП) текстов на естественном языке. Формулы из первого класса, т.е. *l*-формулы, будут называться также К-цепочками.

Каждая формула из множества *Ts*(*B*) представима в виде *d* & *t*, где $d \in Ls(B)$, *t* – тип из *Tps*(*B*). Формулы из множества *Ys*(*B*) являются выражениями вида a_1 & …& a_n & *d*, где $a_1, \ldots, a_n$, $d \in Ls(B)$, *n* имеет разные значения для разных *d*, и цепочка *d* строится из $a_1, \ldots, a_n$ как из элементарных информационных единиц (некоторые из них могут быть немного преобразованы) однократным применением некоторого правила построения.

В данной работе предлагается оригинальная схема подхода к определению трех класов выводимых формул; эта схема заключается в следующем. Будут сформулированы некоторые высказывания *P[0],...,P[10];* они будут интерпретироваться как правила построения семантических представлений (СП) ЕЯ-текстов из элементов первичного информационного универсума *X(B*), переменных из *V(B)* и нескольких специальных символов при условии, что *B* является концептуальным базисом для рассматриваемой области.

Каждое из этих правил фактически задает некоторую операцию на множестве всевозможных наборов, компоненты которых являются СП простых или

составных выражений естественного языка (ЕЯ). Всего 10 операций достаточно для построения формул, отображающих смысл (или структурированные значения) сколь угодно сложных ЕЯ-текстов. Поэтому можно сказать, что система этих правил задает некоторую полную систему квазилингвистических концептуальных операций.

Классы формул *Ls*, *Ts*, *Ys* для произвольного к.б. *B* определяются совместной индукцией правилами *P[0]* , *P[1]*, …, *P[10]* . Для любого к.б. *B* правило *P[0]* задает начальный запас формул.

Определение 1. Обозначим через *P[0]* высказывание "Если $d \in X(B) \cup V(B)$, $t \in Tp(S(B))$, $tp = tp(B)$, $tp(d) = t$, то $d \in L(B)$, и цепочка вида d & t входит в $T^0(B)$". □

Пусть *B* — произвольный к.б., *L(B)* и $T^0(B)$ — наименьшие множества, задаваемые утверждением *P[0]*, $Lnr_0(B) = L(B)$ (обозначение *"Lnr"* расшифровывается как *"L* нумерованное"). Тогда, очевидно, $Lnr_0(B) = X(B) \cup V(B)$, $T^0(B) = \{ b \mid b = d \,\&\, t, d \in X(B) \cup V(B), t \in Tp(S(B)), t = tp(d)\}$.

Таким образом, в соответствии с правилом P[0] информация о типах элементов первичного информационного универсума X(B) и переменных из V(B) отображается в структуре формул из множества T^0(B). □

Пример 1.. Пусть S_1 – с.с., построенная в примере из параграфа 2.8**,** $B_1 = (S_1, Ct_1, Ql_1)$ – к.б., определенный в параграфе 2.8**,** $B=B_1$. Тогда легко увидеть, что выполняются следующие соотношения:

$$чел,\ П.Сомов,\ НПО\_"Радуга",\ Друзья \in Lnr_0(B),$$

$$Персонал,\ Поставщики \in Lnr_0(B);\ чел\ \&\ \uparrow интс * дин.физ.об \in T^0(B),$$

$$П.Сомов\ \&\ интс * дин.физ.об \in T^0(B);$$

$$НПО\_"Радуга"\ \&\ орг * простр.об * интс \in T^0(B);$$

$$Друзья\ \&\ \{(интс * дин.физ.об, \{интс * дин.физ.об\})\} \in T^0(B),$$

$$Персонал\ \&\ \{(орг, \{интс * дин.физ.об\})\ \})\} \in T^0(B),$$

$$Поставщики\ \&\ \{(орг, \{орг\})\} \in T^0(B).$$

Правило *P[1]* предназначено для присоединения информационных единиц, соответствующих словам "некоторый", "каждый" , "какой-нибудь", "все", "несколько", "большинство" (такие информационные единицы в данной работе называются интенсиональными кванторами) к простым или составным

обозначениям понятий. Поэтому правило *P[1]* позволяет строить формальные аналоги выражений: "некоторый человек", "все люди", "большинство людей", "некоторый человек ростом 175 см", "все тридцатилетние люди", “все города Европы”. Примерами *l*-формул (К-цепочек) для *P[1]*, как последнего примененного правила, являются цепочки

*нек чел, все чел *(Возраст, 30/год), все город * (Регион, Европа)* .

Правило *P[2]* предназначено для построения цепочек вида $f(a_1, \ldots, a_n)$, где f – обозначение функции, $n \geq 1$, $a_1, \ldots, a_n$ – *l*-формулы, построенные с применением каких-то правил из списка *P[0]* , *P[1]*, …, *P[10]*. Например, после применения правила на последнем шаге вывода можно получить цепочки
Города(Европа), Колич-элем(Города(Европа)).

Правило *P[3]* позволяет строить цепочки вида $(a_1 \equiv a_2)$, где a_1, a_2 – *l*-формулы, полученные при помощи любых правил из *P[0]*, …, *P[10]*, и a_1, a_2 обозначают сущности, являющиеся однородными в некотором смысле. Примеры К-цепочек для *P[3]* как последнего примененного правила:

$(y_1 \equiv$ *нек город * (Название, ‘Саратов’))*,

(Директор(АО_”Салют”) $\equiv$ *П.Сомов)* .

Правило *P[4]* позволяет строить К-цепочки вида $r(a_1, \ldots, a_n)$, где r – *n*-арное отношение, $n \geq 1$, $a_1, \ldots, a_n$ – К-цепочки, полученные при помощи некоторых правил из *P[0]*, …, *P[10]* . Примеры К-цепочек для *P[4]* : *Принадлеж(Намюр, Города(Бельгия)), Подмнож(Города(Бельгия), Города(Европа)).*

Правило *P[5]* предназначено для построения К-цепочек вида $d : v$, где d – К-цепочка, не включающая v, v – переменная, и выполнены некоторые условия. При помощи правила *P[5]* можно помечать переменными в семантических представлениях текстов на естественном языке: а) описания различных сущностей, встречающихся в тексте (физических объектов, событий, понятий и др.), б) семантические представления предложений или более крупных фрагментов текста, на которые имеется ссылка в любой части текста. Примерами К-цепочек для правила *P[5*], примененного на последнем шаге вывода, являются выражения

все чел : Z1, Меньше(Возраст(П.Сомов), 30/ год) : P1.

Это правило дает возможность строить семантические представления текстов таким образом, чтобы они отражали референтную (ссылочную) структуру текстов. Демонстрирующие это утверждение примеры приведены ниже.

Правило *P[6*] позволяет строить К-цепочки вида $\neg d$, где d – К-цепочка, удовлетворяющая ряду условий. Примеры К-цепочек для *P[6]* :

¬биолог, ¬Принадлеж(Бонн, Города(Бельгия)). Здесь ¬ обозначает связку "не".

При помощи правила *P[7]* можно строить К-цепочки вида $(a_1 \wedge \ldots \wedge a_n)$ или $(a_1 \vee \ldots \vee a_n)$, где $n > 1$, $a_1, \ldots, a_n$ – К-цепочки, обозначающие однородные в некотором смысле сущности. В частности, $a_1, \ldots, a_n$ могут быть семантическими представлениями высказываний, описаниями физических объектов, описаниями множеств, состоящих из объектов одной природы, описаниями понятий. Следующие цепочки являются примерами К-цепочек (или *l*-формул) для *P[7]* :

(Финляндия ∨ Норвегия ∨ Швеция),

(Принадлеж((Намюр ∧ Гент), Города(Бельгия)) ∧ ¬Принадлеж(Бонн,

Города((Финляндия ∨ Норвегия ∨ Швеция))))).

Назначение правила *P[8]* состоит в том, что оно позволяет строить, в частности, К-цепочки вида $c * (r_1,b_1), \ldots, (r_n,b_n)$, где c – информационная единица из первичного универсума X, обозначающая понятие, для $i = 1, \ldots, n$, r_i – функция одного аргумента или бинарное отношение, b_i обозначает возможное значение r_i для объектов, характеризующихся понятием c. Например, если выбрать соответствующим образом первичные информационные единицы, то после применения на последнем шаге вывода правила P[8], можно получить К-цепочки *чел * (Имя, 'Петр')(Фамилия, 'Сомов'), поворот * (Направление, левое).*

Правило *P*[9] дает возможность строить, в частности, К-цепочки вида $\forall v(des)D$ и $\exists v(des)D$, где $\forall$ – квантор всеобщности, $\exists$ – квантор существования, *des* обозначает понятие ("человек", "город", "целое число" и др.) или составные понятия ("целое число, большее 200" и др.). *D* можно интерпретировать как семантическое представление высказывания с переменной *v* о любой сущности, характеризуемой понятием *des*. Примеры К-цепочек для *P[9]* как правила, примененного на заключительном шаге построения формулы:

∀x1(нат.ч.) ∃x2(нат.ч.) Меньше(x1, x2),

∃y(страна ∗(Регион, Европа))Больше(Колич(Города(y)), 15).

Правило *P[10]* позволяет строить, в частности, К-цепочки вида $<a_1,\ldots,a_n>$, где $n > 1$, $a_1,\ldots,a_n$ – К-цепочки. Цепочки, получаемые с использованием правила *P[10]* на последнем шаге вывода, интерпретируются как обозначения *n*-мерных векторов. Компонентами такого вектора могут быть не только обозначения чисел, объектов, но и семантические представления выражений, множеств, понятий и др. Используя правила *P[10]* и *P[4]* , можно построить цепочку

Учиться1(<Агент1, нек чел ∗(Имя, 'Петр')><Учеб.заведение, МГУ>,

<Начал. момент, 1996>), где *Агент1, Учеб.заведение, Начал. момент* – обозначения тематических ролей, т.е. обозначения отношений между значением глагола "учиться" и значениями зависящих от него в предложениях групп слов.

3.2.2. Схема определения трех классов формул, порождаемых концептуальными базисами

Рассмотрим более детально предлагаемую оригинальную схему подхода к определению трех класов выводимых формул.

Определение 2. Если *B* - произвольный концептуальный базис, то пусть

(а) $D(B) = X(B) \cup V(B) \cup \{',' , '(', ')', ':' , '*', '<', '>'\}$,

(б) $Ds(B) = D(B) \cup \{' \&'\}$, (в) $D^+(B)$ и $Ds^+(B)$ — множества всех непустых конечных последовательностей элементов из $D(B)$ и $Ds(B)$, соответственно. □

Если $1 \le i \le 10$, то для любого к.б. *B* и для $k = 1,\ldots,i$ утверждения $P[0],\ldots,P[i]$ определяют совместной индукцией некоторые множества формул $Lnr_i(B) \subset D^+(B)$, $T^0(B)$, $Tnr_i^1(B),\ldots$, $Tnr_i^i(B)$, $Ynr_i^1(B),\ldots$, $Ynr_i^i(B\ ) \ ) \subset Ds^+(B)$. Множество $Lnr_i(B)$ рассматривается как главный подкласс формул, порождаемых правилами $P[0],\ldots,P[i]$. Формулы из этого множества предназначены для описания содержания (смысловых структур) ЕЯ-текстов.

Если $1 \le k \le i$, то множество $Tnr_i^k(B)$ состоит из цепочек вида $b\&t$, где $b \in Lnr_i(B)$, $t \in Tp(S(B))$, и *b* понимается как результат применения правила *P[k]* к некоторым более простым формулам на последнем шаге вывода. Надо добавить, что при построении *b* из элементов *X(B)* и *V(B)* могут использоваться любые

правила *P[0],...,P[k],...,P[i]*; эти правила можно применять произвольно много раз. Если к.б. *B* выбран для описания некоторой области, то *b* можно понимать как СП текста или фрагмент СП текста, относящегося к данной области. В этом случае *t* можно рассматривать как описание вида сущностей, характеризуемых этим СП или фрагментом СП. Кроме того, *t* может квалифицировать *b* как СП повествовательного текста. Номер *i* интерпретируется в этих обозначениях как максимальный номер правила из списка P[0], P[1], … , P[10], которое мы используем для того, чтобы определить множества формул**.**

Таким образом, как будет показано ниже, $Lnr_4(B),..., Lnr_{10}(B)$ включа.ют формулы *Элем(П.Сомов, Друзья(И.Семенов)), Элем(АО_"Салют", Поставщики(НПО_"Радуга")),* и $Tnr_4^4(B),..., Tnr_4^{10}(B)$ включают формулы *Элем(П.Сомов, Друзья(И.Семенов)) & сообщ, Элем(АО_"Салют", Поставщики(НПО_"Радуга")) & сообщ,* где *сообщ* — выделенный сорт $P(B_1)$ ("смысл сообщения").

Каждая цепочка $c \in Ynr_i^k(B)$, где $1 \le k \le i$, может быть представлена в виде $c = a_1 \,\&\, a_2 \,\&\, ... \,\&\, a_m \,\&\, b$, где $a_1,...,a_m$, $b \in Lnr_i(B)$. Кроме того, найдется такое $t \in Tp(S(B))$, что цепочка $b \,\&\, t$ принадлежит $Tnr_i^k(B)$. Цепочки $a_1,...,a_m$ строятся с помощью любых правил из списка *P[0],...,P[10]*, а цепочка b построена из "блоков" $a_1,...,a_m$ (некоторые из них могут быть немного изменены) применением только один раз правила *P[k]*. Возможное количество "блоков" $a_1,...,a_m$ зависит от *k*. Таким образом, множество $Ynr_i^k(B)$ фиксирует результат применения правила *P[k]* один раз. Ниже мы увидим, что множества $Ynr_4^4(B_1),...,$ $Ynr_4^{10}(B_1)$ включают формулы

Элем & П.Сомов & Друзья((И.Семенов) & Элем(П.Сомов, Друзья((И.Семенов)),

Элем & АО_"Салют" & Поставщики(НПО_"Радуга") & Элем(АО_"Салют", Поставщики(НПО_"Радуга")).

Пусть для $i = 1,...,10$ $\quad T_i(B) = T^0(B) \cup Tnr_i^1(B) \cup ... \cup Tnr_i^i(B);$

$Y_i(B) = Ynr_i^1(B) \cup ... \cup Ynr_i^i(B); \quad Form_i(B) = Lnr_i(B) \cup T_i(B) \cup Y_i(B)$.

Будем интерпретировать $Form_i(B)$ как множество формул, порождаемых к.б. В. Это множество представляет собой объединение трех классов формул, главным из которых является $Lnr_i(B)$. Формулы из этих трех классов будем называть соответственно *l*-формулами, *t*-формулами, и *y*-формулами. Класс *t*-

формул необходим для того, чтобы связать тип из $Tp(S(B))$ с каждым $b \in Lnr_i(B)$, где i = 1,...,10. Для $i = 0,...,9$, $Lnr_i(B) \subseteq Lnr_{i+1}(B)$. Множество $Lnr_{10}(B)$ называется стандартным концептуальным языком (стандартным К-языком, СК-языком) в стационарном базисе B и обозначается через $Ls(B)$. Поэтому l-формулы будут часто называться К-цепочками.

Множество $T_{10}(B)$ обозначается через $Ts(B)$. Для любого к.б. B и для любой формулы A из множества $Ts(B)$ существуют такой тип $t \in Tp(S(B))$ и такая формула $C \in Ls(B)$, что $A = C \ \& \ t$. Для построения СП текстов будут использоваться только формулы из $Ls(B)$ и $Ts(B)$, т.е. l-формулы и t-формулы. Y-формулы рассматриваются как вспомогательные и нужны для того, чтобы сформулировать некоторые полезные свойства множеств $Ls(B)$ и $Ts(B)$.

3.3. Использование интенсиональных кванторов в формулах

В параграфе 2.8 было введено понятие *интенсионального квантора*. Этот термин используется для обозначения информационных единиц (другими словами, семантических единиц), соответствующих, в частности, словам и выражениям “каждый”, “некоторый”, “произвольный”, “какой-нибудь”, “определенный”, “все”, “несколько”, “большинство”, “почти все”. Совокупность интенсиональных кванторов делится на два подкласса, обозначаемые через Int_1 и Int_2 . Это осуществляется следующим образом. Компонентом каждого концептуального базиса B вида (2.8.2) является система кванторов и логических связок (с.к.л.с.) Ql вида (2.8.1). Компонентами с.к.л.с. Ql вида являются, в частности, два выделенных сорта int_1 и int_2 . Это дает возможность для $m=1,2$ определить Int_m как $\{x \in X | \ tp(x) = int_m\}$, где первичный информационный универсум X является одним из компонентов концептуально-объектной системы $Ct(B)$ вида (2.7.1)

Элементы множества Int_1 соответствуют значениям выражений “каждый”, “какой-то”, “некоторый”, “произвольный” и т. д. в случаях, когда эти выражения являются частями групп слов, и эти группы связаны с единственным числом. Элементы множества Int_2 интерпретируются как семантические единицы, соответствующие выражениям “все”, “несколько”, “почти все”,

“многие” и т. д. ; минимальное требование к Int_2 заключается в том,, чтобы Int_2 содержало семантическую единицу, соответствующую слову “все”.

Правило P[1] позволяет нам присоединять интенсиональные кванторы к простым или составным обозначениям понятий. В результате применения этого правила, во-первых, строятся: (а) *l*-формулы вида *Int.qr Conc.expr* , где *Int.qr* – интенсиональный квантор из *Int(B)*, а *Conc.expr* – простое или составное обозначение понятия. Во-вторых, строятся *t*-формулы вида *Int.qr Conc.expr & t* , где *t* – тип из множества *Tp(S(B))*.

Например, можно выбрать к.б. *B* так, что в этом базисе с помощью правил P[0] и P[1] можно будет построить *l*-формулы

*нек город, нек город * (Назв, “Чита”),*

каждый город, каждый человек(Квалиф., студент),*

все город, все город(Страна, Россия)*

и *t*-формулы *нек. город & простр. об, все город & {простр. об}* ,

каждый человек(Квалиф, студент) & интс * дин.физ.об* .

Определение 1. Если *B* — произвольный к.б., то для *m* = 1, 2

$Int_m(B) = \{q \in X(B) \mid tp(q) = int_m(B)\}$, $Int(B) = Int_1(B) \cup Int_2(B)$,

Tconc(B) = {t∈ Tp(S(B)) | t начинается с символа ‘↑’ }∪ Spectp ,

где Spectp = {[↑сущн], [↑пон], [↑об]}. □

Напомним, что в параграфе 2.6 выражения *[↑сущн], [↑пон], [↑об]* интерпретируются как такие символы (т.е. неделимые единицы), которые являются информационными единицами, соответствующими словам “сущность” , “понятие” (или “концепт”) и “объект”. В данной работе термин “сущность” является наиболее общим. Объектами называются все те сущности, которые не являются понятиями.

Используя правила P[0] и P[1], мы можем строить *l*-формулы вида *q d*, где *q∈Int(B), d∈ X(B), tp(d) ∈ Tconc(B).* Так, рассматривая к.б. B_1, определенный в параграфе 2.8, мы можем построить *l*-формулы

нек чел, нек тур.гр, нек понятие, все чел, все тур.гр, все понятие.

Можно также строить более сложные цепочки вида *q descr*, где *q* — интенсиональный квантор, *descr* – составное обозначение понятия. Для построения цепочки *descr* используются правила P[0], P[1], правило P[8] (см.

параграф 3.6), и, возможно, некоторые другие правила . Например, для к.б. B_1 будет возможно построить *l*-формулы *нек тур.гр * (Колич, 12), все тур.гр * (Колич, 12).* Эти формулы понимаются как семантические представления (СП) выражений “некоторая туристическая группа из 12 человек” и “все тургруппы из 12 человек”.

Переход от *l*-формулы *c*, обозначающей понятие, к *l*-формуле *q c*, где *q* — интенсиональный квантор, описывается с помощью специальной функции h.

Определение 2. Пусть B — произвольный к.б., S = S(B), Tp(S) – множество типов, порождаемых сортовой системой S . Тогда отображение

h: {1,2} × Tp(S) → Tp(S) задается следующим образом:

(а) если u∈ Tp(S) и цепочка ↑u входит в Tp(S), то h(1, ↑u) = u, h(2, u) = {u} ;

(б) h(1, [↑сущн]) = [сущн], h(1, [↑пон]) = [пон], h(1, [↑об]) =[об],

h(2, [↑сущн]) = {[сущн]}, h(2, [↑пон]) = {[пон]}, h(2, [↑об]) = {[об]} . □

С точки зрения построения СП текстов, отображение h описывает преобразование типов как следующие переходы:

(а) от понятия “человек”, “туристическая группа” к СП выражений “некоторый человек”, “каждый человек”, “любой человек”, “некоторая туристическая группа”, “любая туристическая группа” и. т. д. (в случае, когда первый аргумент h равен 1) и к СП выражений “все люди”, “все туристические группы”, и т. д. . (в случае, когда первый аргумент h равен 2); (б) от выражений “сущность”, “понятие”, “объект” к СП выражений “некоторая сущность”, “произвольная сущность”, “некоторое понятие”, “произвольное понятие”, “некоторый объект”, “произвольный объект”и т. д. (если первый аргумент h равен 1) и к СП выражений “все сущности”, “все понятия”, “все объекты” и т. д. (если первый аргумент h равен 2).□

Определение 3. Через P[1] обозначим высказывание

“Пусть c ∈ L(B) \ V(B), u∈Tconc(B), k∈{0, 8}, и цепочка c & u входит в T^k(B). Пусть m∈ {1, 2}, q∈Int_m, t = h(m,u), и b — цепочка вида q c. Тогда b∈L(B), цепочка вида b & t входит в T^1(B), и цепочка вида q&a&b входит в Y^1(B).” □

Пример 1. Пусть B — к.б. B_1, построенный в параграфе 2.8; *L(B), T^0(B), T^1(B), Y^1(B)* — наименьшие множества, совместно определяемые высказываниями P[0] и P[1]. Тогда легко убедиться в справедливости следующих соотношений:

*чел∈ L(B)\V(B), чел & ↑интс * дин.физ.об∈ T^0(B), нек∈ Int_1(B),*

h(1, ↑интс дин.физ.об) = интс * дин.физ.об ⇒*

*нек чел∈L(B), нек чел & интс * дин.физ.об ∈ T^1(B), нек & чел & нек чел∈ Y^1(B);*

все∈ Int_2(B), h(2, ↑интс дин.физ.об) = {интс* дин.физ.об } ⇒*

*все чел∈L(B), все чел & {интс * дин.физ.об }∈ T^1(B), все&чел&всечел∈ Y^1(B);*

тур.гр∈ L(B)\V(B),тур.гр& ↑{интс дин.физ.об }∈ T^0(B),*

h(1, ↑{интс дин.физ.об }) = {интс* дин.физ.об},*

*h(2, ↑{интс * дин.физ.об }) = {{интс* дин.физ.об }} ⇒*

нек тур.гр, все тур.гр∈L(B), нек тур.гр & {интс}, все тур.гр & {{интс}}}∈ T^1(B),

нек & тур.гр& нек тур.гр∈ Y^1(B); все & тур.гр & все тур.гр∈ Y^1(B);

понятие∈ L(B)\V(B), понятие&[↑пон]∈ T^0(B),

h(1,[↑пон]) = [пон], h(2,[↑пон]) = {[пон]} ⇒ нек понятие, все понятие∈L(B),

нек понятие&[пон], все понятие & {[пон]}∈ T^1(B),

нек & понятие & нек понятие ∈ Y^1(B), все & понятие & все понятие∈ Y^1(B).□

Комментарий к правилу P[1]. Фрагмент правила P[1] “Пусть c ∈ L(B) \ V(B), u∈Tconc(B), k∈{0, 8}, и цепочка c & u входит в T^k(B)” означает, что u – тип понятия (т.к. либо начинается с символа ‘↑’, либо является одним из символов [↑сущн], [↑пон], [↑об]), при k=0 c – простое обозначение понятия (c∈X(B)); при k=8 c – составное обозначение понятия. Такие составные обозначения понятий будут строиться с помощью правила P[8]; примерами таких выражений являются выражения человек* (Область.деят., биология), город*(Страна, Россия)).

Правило P[1] будет очень часто использоваться при построении СП текстов, т.к. оно нужно для построения семантических образов выражений с существительными. Например, пусть T1=”Откуда поступил двухтонный алюминиевый контейнер?”. Тогда в результате выполнения первого шага построения СП T1 можно получить выражение

E1 = *нек контейнер1 * (Вес, 2/Тонна>) (Материал, алюминий)* ,

а после выполнения второго шага – заключительное выражение

E2 = *Вопрос (x1, Ситуация (s1, поступление2 *(Место1, x1) (Объект1,*

*нек контейнер1 * (Вес, 2/Тонна>) (Материал, алюминий)))).*

Более подробно такого рода шаги при построении СП текстов рассматриваются ниже в данной главе и в главе 4.

Часто используемые обозначения. Рассмотрим обозначения, которые в дальнейшем будут часто использоваться в примерах. Для k = 1,...,10 правило P[k] утверждает, что некоторая формула b входит в L(B), некоторая формула b & t принадлежит $T^k(B)$, где $t \in Tp(S(B))$, и некоторая формула z принадлежит $Y^k(B)$. Если $1 \le i \le 10$, B - произвольный к.б., то правила P[0], P[1],... P[i] определяют совместной индукцией множества формул L(B), $T^0(B)$, $T^1(B)$,..., $T^i(B)$, $Y^1(B)$,..., $Y^i(B)$.

Обозначим эти множества через $Lnr_i(B)$, $T^0(B)$, $Tnr_i^1(B)$,... $Tnr_i^i(B)$, $Ynr_i^1(B)$,..., $Ynr_i^i(B$; семейство, состоящее из всех этих множеств, обозначим через $Globset_i(B)$.

Пусть $n \ge 1$, $Z_1,...,Z_n \in Globset_i(B)$, $w_1 \in Z_1,...,w_n \in Z_n$. Тогда, если эти соотношения для формул $w_1,...,w_n$ являются следствием применения некоторых правил $P[l_1],...,P[l_m]$, где $m \ge 1$, то обозначим этот факт выражением вида $B(l_1,..., l_m) \Rightarrow w_1 \in Z_1,...,w_n \in Z_n$. Последовательность $l_1,..., l_m$ может содержать повторяющиеся номера. В выражениях такого рода будет часто пропускаться символ B в обозначениях множеств $Z_1,...,Z_n$; кроме того, будут использоваться выражения $w_1,w_2 \in Z_1$, $w_3,w_4,w_5 \in Z_2$ и т. д. Используя эти обозначения, некоторые соотношения, полученные в Примере 1, можно представить следующим образом:

$B_1(0,1) \Rightarrow$ *все чел, все тур.гр* $\in Lnr_1$,

*все чел & {интс * дин.физ.об}, все тур.гр & {{интс * дин.физ.об}}* $\in Tnr_1^1$,

все & чел & все чел $\in Ynr_1^1$, *все& тур.гр & все тур.гр* $\in Ynr_1^1$.

Выражение $B_1(0,1) \Rightarrow$ *все чел* $\in Lnr_1$, *все чел & {интс * дин.физ.об}* $\in Tnr_1^1$ **равносильно выражению**

$B_1(0,1) \Rightarrow$ *все чел* $\in Lnr_1(B_1)$, *все чел & {интс * дин.физ.об}* $\in Tnr_1^1(B)$. □

Использование реляционных символов и разметка формул

3.4.1. Правила для применения реляционных символов

Правило P[2] позволяет нам, в частности, строить К-цепочки вида $f(a_1,...,a_n)$, где f обозначает функцию с n аргументами $a_1,...,a_n$. Правило P[3] предназначено для построения К-цепочек вида $(a_1 \equiv a_2)$, где a_1 и a_2 обозначают сущности, характеризующиеся типами, сравнимыми друг с другом для отношения конкретизации |—. Используя последовательно P[2] и P[3], мы сможем строить К-цепочки вида $(f(a_1,...,a_n) \equiv b)$, где b — значение f для аргументов $a_1,...,a_n$.

Отметим, что для с.с. S множество главных типов Mtp(S) = Tp(S)\ {[↑сущн], [↑об], [↑пон]} (см. параграф 2.6).

Определение 1. Пусть B — произвольный к.б., S = S(B). Тогда:

(а) $R_1(B)$ = {d∈ X(B) | найдется такой тип t∈ Mtp(S), что t начинается с символа '(' и tp(d) — цепочка вида {t} }; (б) для произвольного n>1, $R_n(B)$ = {d∈ X(B) | найдутся такие $t_1,...,t_n \in$ Mtp(S), что tp(d) - цепочка вида $\{(t_1,...,t_n)\}$ }; (в) для произвольного n>1, $F_n(B) = F(B) \cap R_{n+1}(B)$.

Если n≥1, то элементы множества $R_n(B)$ будем называть n-арными реляционными символами, а элементы $F_n(B)$ будем дополнительно называть n-арными функциональными символами.□

Легко показать, что для произвольного к.б. B и произвольных k,m>1 *из* k≠m *следует* $R_k(B) \cap R_m(B)=\varnothing$.

Определение 2. Обозначим через P[2] высказывание

"Пусть n ≥ 1, f∈ $F_n(B)$, tp = tp(B), $u_1,...,u_n,t \in$ Mtp(S(B)), tp(f) = {($u_1,...,u_n,t$)}; для j = 1,...,n, 0≤ k≤ i, $z_j \in$ Mtp(S(B)), $a_j \in$ L(B), цепочка a_j & z_j принадлежит $T^k_j(B)$; если a_j не входит в V(B), то u_j|— z_j (т. е. z_j является конкретизацией типа u_j); если $a_j \in$ V(B), то u_j и z_j сравнимы для отношения конкретизации |—. Пусть b — цепочка вида $f(a_1,...,a_n)$. Тогда

$b \in L(B), b \,\&\, t \in T^2(B), f \,\&\, a_1 \,\&\, ... \,\&\, a_n \,\&\, b \in Y^2(B)$".□

Перед тем, как сформулировать следующее утверждение, следует напомнить, что символ ‘≡’ является элементом первичного информационного универсума *X(B)* для произвольного к.б. *B*.

Определение 3. Обозначим через P[3] высказывание

“Пусть a_1, $a_2 \in L(B)$, u_1, $u_2 \in Mtp(S(B))$, типы u_1 и u_2 сравнимы для отношения конкретизации |—. Пусть для m = 1, 2, $0 \le k[m] \le i$, a_m & $u_m \in T^{k[m]}(B)$; P - сорт “смысл сообщения” для к.б. B, b — цепочка $(a_1 \equiv a_2)$.Тогда

$b \in L(B)$, $b\&P \in T^3(B)$, и цепочка a_1 & ≡ & a_2 & b входит в множество $Y^3(B)$.” □

В правилах P[2] и P[3] символ i обозначает неизвестное число, такое что $2 \le i \le 10$. Дело в том, что правила P[0] – P[3] и последующие правила будут использоваться вместе с объединяющим их определением, которое начинается с такой фразы: «Пусть B – произвольный к.б., $1 \le i \le 10$.» Число i будет интерпретироваться как максимальный номер правила, которое будет использоваться для построения формулы. Например, если I = 3, то мы можем использовать правила с номерами 0, 1, 2, 3, но не можем использовать правила с номерами 4 - 10. Параметр i в правилах вывода позволяет нам после введения очередного правила с номером i+1 определить формальный язык Lnr_{i+1} (B) и исследовать выразительные возможности этго языка.

Пример 1. Пусть B_1 — к.б., построенный в параграфе 2.8; $i = 3$;

b_1 = *Поставщики(НПО_”Радуга”)* , b_2 = *Колич (Поставщики(НПО_”Радуга”)* ,

b_3 = *(Колич (Поставщики(НПО_”Радуга”) ≡12)* , b_4 = *Колич(все понятие)*,

b_5 = *Колич(все химик)*, b_6 = *(все химик ≡x1)* , b_7 = *Колич(x1)* .

Тогда легко убедиться в справедливости следующих соотношений (принимая во внимание обозначения, введенные в конце предыдущего параграфа):

$B_1 \Rightarrow$ *Поставщики &{(орг, {орг})}* $\in T^0$,

*НПО_”Радуга” & орг * простр.об * интс, Колич&{({[сущн]}, нат)}* $\in T^0$;

$B_1(1, 2) \Rightarrow b_1$ *& {орг}* $\in Tnr_3^2$;

$B_1(0, 2, 2) \Rightarrow b_2 \in Lnr_3$, b_2 *& нат* $\in Tnr_3^2$, *Колич &* b_1 *&* $b_2 \in Ynr_3^2$;

$B_1(0, 2, 2, 0, 3) \Rightarrow b_3 \in Lnr_3$, b_3 *& сообщ* $\in Tnr_3^3$;

$B_1(0, 2, 2) \Rightarrow b_4, b_5 \in Lnr_3$, b_4 *& нат*, b_5 *& нат* $\in Tnr_3^2$;

$B_1(0, 1, 3) \Rightarrow b_6 \in Lnr_3$, b_6*&сообщ* $\in Tnr_3^3$;

x1∈V(B₁), tp(x1) = {[сущн]} , B₁(0, 2) ⇒ b₇∈Lnr₃, b₇ & нат∈ Tnr₃².□

Определение 4. Обозначим через P[4] высказывание

"Пусть n≥1, r∈ $R_n(B)\backslash F(B)$, $u_1,...,u_n \in Mtp(S(B))$, tp = tp(B), tp(r) — *цепочка* {($u_1,...,u_n$)} *при* n > 1 *или* { u_1} *при* n = 1*; для* j = 1,...,n, 0 ≤ k[j] ≤ i, $z_j \in Mtp(S(B))$, $a_j \in L(B)$, *цепочка* a_j & z_j *принадлежит* $T^{k[j]}_j(B)$; *если* $a_j \notin V(B)$*, то* u_j|— z_j*; если* $a_j \in V(B)$, *то* u_j *и* z_j *сравнимы для отношения конкретизации* |—. *Пусть* b — *цепочка вида* $r(a_1,...,a_n)$, P = P(B) — *сорт "смысл сообщения" для* B. *Тогда*

b∈ L(B), b & P∈ $T^4(B)$, и r& a_1&...& a_n & b∈ $Y^4(B)$".□

Пример 2. Пусть B_1 - к.б., рассмотренный в параграфе 2.8; *i =4*;

b_8 = *Меньше(10000, Колич(все химик)),*

b_9 = *Меньше(5000, Колич(все понятие)),*

b_{10} = *Элем(биолог, все понятие),*

b_{11} = *Элем(АО_"Старт", Поставщики(НПО_"Радуга"))* ,

b_{12} = *(П.Сомов ≡ Директор (АО_ "Старт"))* ,

b_{13} = *Знает(П.Сомов, (Колич(Поставщики(НПО_"Радуга"))≡12)),*

b_{14} = *Знает(И.Семенов, (П.Сомов ≡Директор (АО_ "Старт")))* ,

b_{15} = *Меньше(10000, Колич(x1)).*

Принимая во внимание определение к. о. с. $Ct(B_1)$ (см. пример в параграфе 2.7) и применяя правила P[0],...,P[4], мы получаем следующие соотношения:

$B_1(0, 1, 2, 3, 4) ⇒ b_8,..., b_{15} ∈ Lnr_4$, для m = 8,…, 15 b_m & сообщ∈ Tnr_4^4,

Меньше & 1000 & Колич(все химик) & Меньше (10000, Колич(все химик)) ∈ Ynr_4^4.

3.4.2. Правило, позволяющее помечать формулы

Основное назначение правила P[5] заключается в том, чтобы с помощью переменных из множества V(B) , где B – рассматриваемый к.б., помечать в семантических представлениях (СП) ЕЯ-текстов: (а) описания различных сущностей, упомянутых в тексте (физических объектов, событий, понятий и т. д.), (б) фрагменты, являющиеся семантическими представлениями предложений и более крупных частей текстов, на которые имеются ссылки в любой части

текста. С помощью этого правила (в сочетании с другими правилами) окажется возможным отражать референтную (ссылочную) структуру дискурсов, в которых имеются, в частности, выражения (а) этот прибор, на этом заводе, в этом городе, ему , о нем , (б) данный метод, это распоряжение, эту команду , этот вопрос, об этом, про это.

Определение. Обозначим через P[5] высказывание

"Пусть $a \in L(B) \setminus V(B)$, $0 \le k \le i$, $k \ne 5$, $t \in Mtp(S(B))$, $a\& t \in T^k(B)$; $v \in V(B)$, $z \in Mtp(S(B))$, $v \& z \in T^0(B)$, $z \text{ |— } t$, v не является подцепочкой цепочки a. *Пусть* b - цепочка вида a : v. Тогда $b \in L(B)$, $b\& t \in T^5(B)$, $a\& v\& b \in Y^5(B)$".□

Условие "k≠5 " вводится для того, чтобы правило P[5] нельзя было применять произвольное число раз подряд; в противном случае мы могли бы получать формулы с "избыточной разметкой", то есть выражения вида $a : v_1 : v_2 \ldots v_n$, где $n > 1$, и $v_1, v_2, \ldots, v_n \in V(B)$.

Пример 3. Рассмотрим, как м прежде, к.б. B_1, построенный в параграфе 2.8. Пусть i = 5, $a_1 = b_3 =$ *(Колич (Поставщики(НПО_"Радуга") ≡ 12),* $k_1 = 3$, $t_1 =$ *сообщ = P(*B_1*).* Тогда, очевидно, $a_1\& t_1 \in Tnr_5^3(B_1)$.

Предположим, что v_1 = P1, z_1 = сообщ = P(B_1). Тогда, в соответствии с определением базиса B_1, $v_1\& z_1 \in T^0(B_1)$; кроме того, $z_1 \text{|—} t_1$ (т. к. $z_1 = t_1$), и v_1 не является подстрокой a_1.

Пусть $b_{16} =$ *(Колич (Поставщики(НПО_"Радуга")) ≡ 12) : P1* . Тогда, в соответствии с правилом P[5], $b_{16} \in Lnr_5(B_1)$, b_{16} *&* *сообщ* $\in Tnr_5^5(B_1)$, $a_1\& v_1\& b_{16} \in Ynr_5^5(B_1)$.

Пусть $b_{17} =$ *Поставщики(НПО_"Радуга"): x3* . Тогда легко видеть, что

$$B_1(0, 2, 5) \Rightarrow b_{17} \in Lnr_5, b_{17} \&\{орг\} \in Tnr_5^5,$$

$$Поставщики(НПО\_"Радуга") \& x3 \& b_{17} \in Ynr_5^5.$$

В выражении b_{16} переменная P1 помечает СП фразы П1 = "У НПО "Радуга" имеется 12 поставщиков". Поэтому, если это выражение является частью длинной цепочки, то справа от вхождения b_{16} в такую цепочку можно использовать переменную P1 для повторного представления смысла указанной фразы П1 вместо значительно более длинного СП фразы П1. В цепочке b_{17}

переменная x3 является меткой множества, состоящего из всех поставщиков научно-производственного объединения “Радуга”.

3.5. Использование логических связок “не”, “и” , “или”

Правила P[6] и P[7] в сочетании с другими правилами позволяют по сравнению с языком логики предикатов более полно моделировать (на уровне семантических представлений текстов) способы использования связок “не”, “и”, “или” в предложениях на русском, английском и многих других языках. В частности, конструктивно учитывается существование фраз вида “Этот препарат выпускается не в Польше”, “Профессор Сухов работает не в МГУ”, “Этот патент внедрен в Австрии, Венгрии, Нидерландах и Великобритании”. С этой целью, во-первых, допускается присоединение связки $\neg$ (“не“) не только к семантическим представлениям (СП) фраз, обозначающих высказывания, но и к обозначениям предметов, событий, понятий. Во-вторых, разрешается соединять связками $\wedge$ (“конъюнкция”, т.е. логическое “и“) и $\vee$ (“дизъюнкция”, т.е. логическое “или”) не только СП высказываний, но и обозначения предметов, ситуций, понятий, множеств предметов и т.д.

Правило P[6] предназначено для построения l-формул вида $\neg$a, где a$\in$ Lnr_i(B).

Определение 1. Обозначим через P[6] высказывание «Пусть a$\in$ L(B), t$\in$ Mtp(S(B)), $0 \le k \le i$, k$\notin$ {2,5,10}, a& t$\in$ T^6(B), b — цепочка вида $\neg$ a. Тогда b$\in$ L(B), b & t$\in$ T^6(B), цепочка вида $\neg$ & a& b входит в множество Y^6(B).” □

Замечание 1. Условие k$\notin$ {2,5,10} означает следующее: если l-формула a выведена каким-либо образом с помощью правил P[0], P[1],...,P[i], то правило P[2] или P[5] или P[10] не может быть применено на последнем шаге вывода. Таким образом, условие k$\neq$2 означает, что не разрешается строить l-формулы вида $\neg f(d_1,\ldots,d_m)$, где f$\in$ F(B). Условие k$\neq$5 вводится для того, чтобы по любому выражению вида $\neg$ a:v, где v$\in$ V(B), можно было однозначно найти то правило, которое применялось последним; этим правилом будет P[5]. Условие k$\neq$10 запрещает строить выражения вида $\neg$ $<C_1, \ldots, C_m>$, т.е. запрещает присоединять связку “не“ к обозначениям упорядоченных наборов.

Пример 1. Пусть B_1 – к.б., построенный в параграфе 2.8, i=6. Тогда легко видеть, что выполняются следующие соотношения:

B_1 *(0, 6) => ¬ биолог* $\in Lnr_6$, *¬ биолог & ↑интс*дин.физ.об* $\in Tnr^6_6$.

B_1*(0,6,4,4) => Знает (П.Сомов, Сейчас, Явл1(И. Семенов, ¬ биолог))* $\in Tnr^4_6$;

B_1*(0,4,4,6) => ¬ Знает(П.Сомов, Сейчас, Явл1(И.Семенов, химик))* $\in Tnr^6_6$.

Правило P[7] позволяет строить *l*-формулы вида ($a_1 \wedge a_2 \wedge \ldots \wedge a_n$) и вида ($a_1 \vee a_2 \vee \ldots \vee a_n$). Например, формулы *(химик ∨ биолог), (математик ∧ художник), (Имя (x1, 'Сергей') ∧ Фам(x1, 'Жаворонков') ∧ Квалиф(x1, химик)).*

Определение 2. Обозначим через P[7] высказывание: "Пусть n>1, t∈ Mtp(S(B)), для m=1,…,n, $0 \le k[m] \le i$, a_m & t∈ $T^{k[m]}(B)$, s∈ {∧, ∨}, b – цепочка вида (a_1 s a_2 s … s a_n). Тогда *b∈ L(B), b & t∈ $T^7(B)$, s & a_1&…& a_n & b∈$Y^7(B)$"*.

Замечание 2. Данное правило требует, чтобы все выражения, соединенные за один шаг логической связкой, имели один и тот же тип. Так как *t* не обязательно является сортом «смысл сообщения», то по правилу P[7] можно соединять логическими связками не только семантические представления высказываний, но и обозначения различных объектов, простые и составные обозначения понятий, простые и составные обозначения целей интеллектуальных систем.

Пример 2. Пусть B_1 – к.б., построенный в в параграфе 2.8, i= 7 ,

b_1 *= (А.Зубов ∧ И.Семенов),* b_2 *= (химик ∨ биолог),*

b_3 *= ((Колич (Друзья (А.Зубов)) ≡3) :* P_1 *∧ Знает (П.Сомов, Сейчас,* P_1*)*

∧ ¬Знает (П.Сомов, Сейчас, Явл (А.Зубов, (химик ∨ биолог)))),

b_4 *= Элем ((АО_"Салют" ∧ АО_"Старт"), Поставщики (НПО_"Радуга"))* .

Тогда легко показать, что B_1*(0,7) =>* b_1, $b_2 \in Lnr_7$, b_1*& интс*дин.физ.об* $\in Tnr^7_7$,

b_2 *&* ↑*интс*дин.физ.об* $\in Tnr^7_7$, B_1*(0,2,2,3,5,4,7,4,4,6,7) =>* $b_3 \in Lnr_7$,

b_3 *& сообщ*$\in Tnr^7_7$; B_1*(0,7,0,2,4) =>* $b_4 \in Lnr_7$, b_4 *& сообщ*$\in Tnr^4_7$.

3.6. Построение составных обозначений понятий и объектов

Правило для построения составных обозначений понятий

Рассмотрим правило P[8], предназначенное для построения составных обозначений понятий. С помощью этого правила строятся *l*-формулы вида $a^*(r_1, d_1)\ldots(r_n, d_n)$ и *t*-формулы вида $a^*(r_1, d_1)\ldots(r_n, d_n) \,\&\, t$, где a – элемент первичного информационного универсума $X(B)$ и интерпретируется как простое обозначение понятия, $n \geq 1$, для $i=1,\ldots,n$, $r_i \in R_2(B)$, d_i – обозначение некоторой сущности. Например, выбирая подходящий конептуальный базис, можно построить *l*-формулы

*город * (Страна, Россия)* , *учебник * (Область, биология), понятие*(Имя.пон, "молекула")* , *туристич.группа*(Количество, 12)(Состав, (химик ∧ биолог))*

и t-формулу *город*(Страна, Россия)* & *↑простр.об* .

Используя правило P[8] вместе с правилом P[1] и другими правилами, можно будет строить составные обозначения объектов и множеств объектов в виде *q des*, где *q* — интенсиональный квантор (см. параграф 1.8), *des* — составное обозначение понятия, построенное с помощью правила *P[8]* на последнем шаге вывода. В частности, формулы

*некотор учебник * (Область, биология), все город * (Страна, Россия) , некотор чел*(Возраст, 18/год>), некотор понятие*(Имя.пон, "молекула"), некотор туристич.группа* (Количество, 12))(Качеств-состав, (химик ∧ биолог))* .

Определение 1. Для произвольного к.б. B $Tconc(B) = \{t \in Tp(S(B)) \mid t$ начинается с $\uparrow\} \cup Spectp$, где $Spectp = \{[\uparrow сущн], [\uparrow пон] , [\uparrow об]\}$.

Поясним обозначения, используемые ниже в определении правила P[8]. Каждый такой элемент c первичного информационного универсума $X(B)$, что $tp(c) \in Tconc(B)$, будем интерпретировать как обозначение понятия. Множество $R_2(B)$ состоит из бинарных реляционных символов (некоторые из них могут соответствовать функциям с одним аргументом); $F(B)$ — множество функциональных символов. Элемент $ref = ref(B)$ из $X(B)$ называется *квантором референтности* (см. параграф 2.8) и интерпретируется как информационная

единица (другими словами, семантическая единица), соответствующая значению слова “некоторый” в выражениях в единственном числе (“некоторая книга”, “некоторая страна” и т. д.); в рассматриваемых примерах *ref(B)* – это символ *нек* ; *P(B)* — обозначение сорта “смысл сообщения” к.б. *B*.

Определение 2. Обозначим через P[8] высказывание

“Пусть $a \in X(B)$, $tp = tp(B)$, $t \in Tconc(B)$, $t = tp(a)$, $P = P(B)$, $ref = ref(B)$. Пусть $n \geq 1$, $\forall m=1,\ldots,n$ $r_m \in R_2(B)$, c_m – цепочка вида *ref a*, $d_m \in L(B)$, h_m – цепочка вида $(r_m(c_m) \equiv d_m)$ в случае $r_m \in R_2(B) \cap F(B)$, и h_m – цепочка вида $r_m(c_m, d_m)$ в случае $r_m \in R_2(B) \setminus F(B)$; если $r_m \in F(B)$, то $h_m \,\&P \in T^3(B)$; если $r_m \notin F(B)$, то $h_m \,\&P \in T^4(B)$. Пусть b - цепочка вида $a^*(r_1,d_1)\ldots(r_n,d_n)$. Тогда $b \in L(B)$, $b\,\& t \in T^8(B)$, $a\,\& h_1 \&\ldots\& h_n \& b \in Y^8(B)$.” □

Пример 1. Предположим, что B_1 — к.б., определенный в параграфе 2.8, $i = 8$. Тогда рассмотрим возможный путь построения формулs b_1 (задаваемой ниже), соответствующей понятию “туристическая группа, состоящая из 12 человек”. При этом будем использовать обозначения, введенные в конце параграфа 3.3.

Пусть a = *тур.гр*, $t = tp(a) = \uparrow\{$*интс * дин.физ.об*$\}$, P = *сообщ* , ref = *нек*,

$$n = 1,\ r_1 = Колич,\ c_1 = ref\ a = нек\ тур.гр,\ d_1 = 12,$$

$$h_1 = (r(c) \equiv d) = (Колич(нек\ тур.гр) \equiv 12).$$

Тогда $B_1(0, 1, 2, 3) \Rightarrow h_1 \in Lnr_3(B_1)$, $h_1\,\&$ *сообщ* $\in Tnr_3^{\,8}(B_1)$.

Пусть $b_1 = a^*(r_1, d_1) =$ *тур.гр* (Колич, 12)*. Тогда из *P[8]* следует, что

$$b_1 \in Lnr_8(B_1),\ b_1\,\& \uparrow\{инс^*\ дин.физ.об\} \in Tnr_8^{\,8}(B_1).\ .\square$$

3.6..2. Построение составных обозначений объектов

Правило P[8] можно использовать для построения составных обозначений разных предметов, ситуаций и множеств предметов или ситуаций. Для построения составного обозначения предмета после правила P[8] применяется правило P[1] и строится выражение вида $ref\ a^*(r_1,d_1)\ldots(r_n,d_n)$. Например, таким образом можно построить выражения *нек город (Страна, Россия) (Колич. жит., 350000), нек чел* (Фам, “Сомов”)(Имя, “Петр”).*

Пример 2. Пусть b_2 = *нек биолог*(Элем, нек тур.гр*(Колич,12))*. Тогда

$B_1(0, 1, 2, 3, 8, 1, 1, 4, 8, 1) \Rightarrow b_2 \in Lnr_8(B_1)$, $b_2\,\&$ *интс* дин.физ.об* $\in Tnr_8^{\,1}(B_1)$.

Цепочку b_2 будем интерпретировать как составное обозначение какого-то (вполне определенного) человека, являющегося биологом и входящего в состав тургруппы из 12 человек. Подцепочку *нек тур.гр*(Колич,12))* будем интерпретировать как обозначение какой-то конкретной (в контексте ситуации общения) тургруппы, состоящей из 12 человек□

Пример 3. Исходя из предположений Примера 1, рассмотрим путь построения возможного СП фразы "А.Зубов включил И.Семенова в туристическую группу, состоящую из 12 человек". Пусть переменная *x1* обозначает момент времени, и b_3 = *(Включ1(А.Зубов , И.Семенов , x1, нек тур.гр*(Колич,12)) ∧Раньше (x1,Сейчас)).*

Тогда $B_1(0, 1, 2, 3, 8, 1, 4, 4, 7) \Rightarrow b \in Lnr_8(B_1)$, b_3 & *сообщ* $\in Tnr_8^7(B_1)$.

Пример 4. Пусть Т1 = "П.Сомов знает, что И.Семенов является директором фирмы, персонал которой включает 38 человек". Это предложение включает составное обозначение фирмы с персоналом из 38 человек. Пусть

b_4 = *Знает (П.Сомов, Сейчас, ((И.Семенов ≡Директор (нек фирма * (Описание, P1): x1)) ∧ (P1 ≡(Колич(Персонал (x1)) ≡38))))* .

Проследим путь вывода формулы b_4 с помощью правил Р[0], Р[1], …, Р[8] . Пусть $i=8$, B_1 – к.б., построенный в параграфе 2.8. Рассмотрим новый к.б. B_2, отличающийся от базиса B_1 тем, что $X(B_2) = X(B1) \cup \{$*фирма, Описание*$\}$, *tp(фирма) = ↑орг * простр.об * интс , tp(Описание)={([ob],P)}.*

Если *a=фирма, t=орг * простр.об * интс* , то, очевидно, *a* & $t \in T^0(B_2)$. По определению к.б. B_2, множество переменных $V(B_2)$ включает такую переменную *P1*, что $tp(P1)=P(B_2)$ = *сообщ* . Пусть

$n=1$, r_1= *Описание*, c_1= *ref a= нек фирма*,

$h_1 = r_1(c_1,d_1)$ = *Описание(нек фирма, P1)*, $b_5 = a^*(r_1,d_1)$= *фирма*(Описание,*P_1*).*

Тогда из правила Р[8] и правил Р[0], Р[1], Р[4] следует, что

$b_5 \in Lnr_i(B_2)$, b_1& *↑орг* простр.об * интс* $\in Tnr_i^8(B_2)$.

Пусть b_6 = *ref* b_1 = *нек фирма * (Описание,*P_1*).* Тогда

$B_2(0,1,4,8,1) \Rightarrow b_6 \in Lnr_i(B)$, b_6 & *орг* простр.об * интс* $\in Tnr_i^1(B_2)$.

Пусть $b_7 = b_6$ *: x1= нек фирма*(Описание,* P_1*): x1.* Тогда

$B_2(0,1, 4, 8,1,5) \Rightarrow b_7 \in Lnr_i$, b_7 & *орг* простр.об * интс* $\in Tnr_i^5$.

Пусть b_8= *(Р1≡(Колич(Персонал(x1))≡38))*. Тогда $B_2(0,2,2,3,3) \Rightarrow b_8 \in Lnr_i$, b_8 & *сообщ* $\in Tnr_i^3$. Пусть b_9 = *(И.Семенов≡Директор(b_7))*, тогда

$$B_2(0,1,4,8,1,5,2,3) \Rightarrow b5 \in Lnr_i,\ b5 \& сообщ \in Tnr_i^3.$$

Пусть b_{10} = *Знает (П.Сомов, Сейчас, ($b_9 \wedge b_8$))*. Тогда

$$B_2(0,1,4,8,1,5,2,2,3,7,4) \Rightarrow b_{10} \in Lnr_i,\ b_{10} \& сообщ \in Tnr_i^4.$$

Легко видеть, что b_{10} совпадает *с* b_4. Значит, мы построили вывод формулы b_4. Рассмотренный метод построения СП текста Т1 является весьма общим; этот метод может использоваться в самых разнообразных случаях для построения СП предложений со сложными причастными оборотами и придаточными определительными предложениями.

3.7. Использование в формулах кванторов существования и всеобщности. Построение обозначений упорядоченных наборов

3.7.1. Применение кванторов существования и всеобщности

Правило Р[9] позволяет строить формулы с кванторами ∃,∀, похожие на формулы логики предикатов первого порядка. Отличие, в частности, заключается в том, что явным образом ограничивается область действия кванторов, и переменные могут обозначать не только предметы, числа, но и множества различных сущностей. С помощью Р[9] можно построить l-формулы вида *Q v (concept) A (v)*, где $Q \in \{\exists, \forall\}$,$v \in V(B)$ – переменная; *concept* – это простое обозначение понятия (в этом случае *concept* – такой элемент первичного информационного универсума *X(B)*, что *tp(concept)* начинается с символа ↑) или составное обозначение понятия; например, *concept* = *страна*(Место, Европа))*, *A(v)* – формула, включающая v и интерпретируемая как СП высказывания.

Пример 1. Выбирая подходящий к.б. В, с помощью Р[9] и нескольких других правил можно построить СП предложения “В каждой стране Европы есть город с количеством жителей, превышающим 30 тысяч человек” следующим образом:
∀x1(страна*(Место,Европа)) ∃x2(город) (Место(x2, x1) ∧ Меньше (30000, Колич.элем. (Жители(x2)))) .

Определение 1. Через P[9] обозначим высказывание

“Пусть $Q\in\{\exists,\forall\}$, $A\in L(B)$, $P=P(B)$, $k\in\{3,4,6,7,9\}$, A & $P\in T^k(B)$, $v\in V(B)$, $tp(v)=[сущн]$ – базовый тип «сущность», A включает символ v, $m\in\{0,8\}$, $concept\in L(B)\backslash V(B)$, $u\in Tconc(B)$, где $Tconc(B)$ – множество всех типов из $Tp(S(B))$, начинающихся с символа $\uparrow$; цепочка вида $concept$ & u входит в множество $T^m(B)$, цепочка A не включает подцепочек видов : v, ∀v, ∃v, и цепочка A не имеет окончания вида :z, где z – произвольная переменная из $V(B)$.

Пусть b=Q v (concept) A. Тогда

$b\in L(B)$, b & $P\in T^9(B)$, Q & v & $concept$ & A& $b\in Y^9(B)$” .

Условие “$k\in\{3,4,6,7,9\}$” означает, что перед использованием правила P[9] (т.е. перед присоединением к формуле квантора существования или всеобщности) должно применяться одно из правил P[3], P[4], P[6], P[7] или P[9].

Пример 2. Пусть $i=9$ и существует такой к.б. B, что выполняются следующие соотношения:

$$город, страна, Европа, Место, Колич.элем., Жители, Меньше \in X(B),$$

$$простр.об, нат.чис, дин.физ.об, интс\in St(B),$$

$$tp(город)= tp(страна)= \uparrow простр.об,\ tp(Европа)=простр.об,$$

$$tp(Место)=\{(простр.об, простр.об)\},$$

$$tp(Колич.элем.)=\{(\{сущн]\}, нат.чис)\},$$

$$tp(Жители)=\{(простр.об, \{дин.физ.об * интс\})\},$$

$$tp(Меньше)=\{(нат.чис, нат.чис)\}, Колич.элем, Жители \in F(B) ,$$

$$x1, x2\in V(B) ,\ \ tp(x1)=tp(x2)=[сущн],\ 30000\in X(B) ,$$

$$tp(30000)=нат.чис, сообщ = P(B).$$

Перечисленные информационные единицы интерпретируются следующим образом: *простр.об, нат.чис, дин.физ.об, интс* – сорта “пространственный объект”, “натуральное число” , “динамический физический объект”, “интеллектуальная система”; *город, страна* – обозначения одноименных понятий; *Европа* – обозначение региона Европа; *Место* - обозначение бинарного отношения, связывающего пространственные объекты; *Колич.элем* – обозначение функции “Количество элементов множества”; *Жители* - обозначение функции, ставящей в соответствие населенному пункту множество

всех его жителей; *Меньше* - обозначение бинарного отношения “меньше” на множестве натуральных чисел.

Пусть $Q_1=\exists$, $v_1=x2$, $concept_1=$*город*, $A_1=$(*Место*(x2, x1)$\wedge$ *Меньше*(30000, *Колич.элем.* (*Жители*(x_2))) , $b_1=Q_1v_1(concept_1)A_1$. Тогда нетрудно проверить, что выполняется следующее соотношение: $B(0,2,2,3,4,4,7,9)\Rightarrow b_1\in Lnr_i$, b_1 & *сообщ*$\in Tnr_i^9$. Пусть $q_2=\forall$, $v_2=x1$, $concept_2=$ *страна* * (*Место*,*Европа*) , $A_2=b_1$, $b_2=Q_2v_2(concept_2)A_2$.Тогда $B(0,2,2,3,4,4,7,9,0,1,4,8,9)\Rightarrow b_2\in Lnr_{i=9}$, b_2 & *сообщ*$\in Tnr_i^9$, Q_2 &v_2 & $concept_2$ & b_1& $b_2\in Y_i^9(B)$.

3.7.2. Построение обозначений упорядоченных наборов

Правило P[10] предназначено для построения l-формул вида $<a_1,\ldots,a_n>$, где $n>1$ и $a_1,\ldots,a_n$ – обозначения некоторых сущностей. Такие формулы будут интерпретироваться как обозначения упорядоченных наборов.

Определение 2. Через P[10] обозначим высказывание:

“Пусть $n>1$, для $m=1,\ldots, n$ выполняются соотношения $a_m\in L(B)$, $u_m\in Tp(S(B))$, $0\le k[m]\le 10$, a_m & $u_m\in T^{k[m]}(B)$. Пусть t - цепочка вида $(u_1,u_2,\ldots,u_n)$, b - цепочка вида $<a_1,\ldots,a_n>$. Тогда

$$b\in L(B),\ b\ \&\ t\in T^{10}(B),\ a_1\&a_2\&\ldots\&\ a_n\ \&\ b\in Y^{10}(B).”$$

Пример 3. Пусть B_1 – к.б., построенный в параграфе 2.8, $i=10$, b_3 - цепочка (*Элем*(x3 ,S1) $\equiv$((x3 $\equiv$<*нек вещ* : x1, *нек вещ* : x2>) $\wedge$ (*Меньше* (x1 , x2)$\vee$(x1$\equiv$ x2)))) . Тогда *S1* можно интерпретировать как обозначение отношения «≤» на множестве вещественных чисел. То есть S_1 обозначает множество всех таких пар (x1, x2), что x1, x2 – вещественные числа, и x1 ≤ x2 . Легко проверить, что $B(0,4,1,5,1,5,4,10,3,7,7,3)\Rightarrow b_3\in Lnr_i(B_1)$, b_3 & *сообщ*$\in Tnr_i^{10}(B_1)$.

3.7.3. Сводная таблица правил P[0] – P[10]

Суммарный объем определений правил P[0] – P[10] и поясняющих их примеров довольно велик. При построении семантических представлений (СП) не только связных текстов, но и большинства отдельных предложений обычно используется значительная часть этих правил, причем в самых разнообразных комбинациях (анализу возможных применений этих правил посвящена Глава 4). В связи с этим представляется целесообразным дать сжатую, недетализированную характеристику каждого правила из списка P[0] – P[10] в приводимой ниже сводной таблице. Эта таблица облегчит анализ использования правил P[0] – P[10] при построении СП ЕЯ-текстов и при формировании фрагментов знаний о мире в примерах этой и последующих глав.

Правило	Результаты применения
P[0]	Начальный запас формул, определяемый первичным информационым универсумом *X(B)* , множеством переменных V(B) и отображением *tp* , задающим типы элементов из этих множеств
P[1]	*l* – формулы вида *q a* или вида *q a* $*(r_1,d_1)\ldots(r_n,d_n)$, где *q* – интенсиональный квантор, *a* – простое обозначение понятия, $1 \le n$, $r_1 \ldots r_n$ – характеристики сущностей
P[2]	*l*– формула вида $f(a_1,\ldots, a_n)$, $1 \le n$, где *f* – имя функции; *t* - формулы вида $f(a_1,\ldots, a_n)$ & *t*, где *t* – тип значения функции f для аргументов $a_1,\ldots, a_n$
P[3]	*l* – формулы вида $(a_1 \equiv a_2)$ и *t* – формулы вида $(a_1 \equiv a_2)$ & *P*, где *P* – сорт «смысл сообщения»
P[4]	*l* – формулы вида $r(a_1,\ldots, a_n)$ и *t* – формулы вида $r(a_1,\ldots, a_n)$ & *P*, где $n \ge 1$, *r* - n-арный реляционный символ
P[5]	*l* - формула вида *form : v*, где *v* – метка формулы *form*
P[6]	по *l* – формуле *form* строится *l* – формула $\neg$ *form* (отрицание)
P[7]	по логической связке $s \in \{\wedge, \vee\}$ и *l* -формулам $a_1,\ldots, a_n$ строится *l*-формула (a_1 s a_2 s … a_n), где $n>1$
P[8]	по простому обозначению понятия *conc*, характеристикам объектов $r_1,\ldots, r_n$ $(n \ge 1)$, *l* -формулам $d_1 \ldots d_n$ строится

	l-формула *conc* $*(r_1,d_1)\ldots(r_n,d_n)$ и *t*-формула *conc* $*(r_1,d_1)\ldots(r_n,d_n)$: *t* , где *t* начинается с символа $\uparrow$. Такие формулы интерпретируются как составные обозначения понятия. Пример: *страна* $*$ *(Место, Европа)* & $\uparrow$*простр. объект*
P[9]	строятся *l*-формулы вида *Qv(conc)A* и *t*-формулы *Qv(conc)A&P*, где Q$\in\{\exists,\forall\}$, v$\in$ V(B) , *conc* - простое или составное обозначение понятия, *A* - *l*-формула, обозначающая высказывание, *P* – сорт «смысл сообщения»
P[10]	Для построения *l*-формул вида $<a_1,\ldots a_n>$, где $n>1$, интерпретируемых как обозначения упорядоченных наборов

Табл. 3.1. Краткая характеристика правил P[1] – P[10].

3.8. Стандартные К-языки. Математическое исследование их свойств

Определение 1. Пусть *B* - произвольный концептуальный базис, тогда:

(а) $D(B) = X(B) \cup V(B) \cup \{',' , '(', ')', ':' , '*', '<', '>'\}$,

(б) $Ds(B) = D(B) \cup \{' \&'\}$, (в) $D^+(B)$ и $Ds^+(B)$ — множества всех непустых конечных последовательностей элементов из *D(B)* и *Ds(B)* соответственно. □

Таким образом, $Ds(B) = X(B) \cup V(B) \cup \{',', '(', ')', ':' , '*', '<', '>', '\&' \}$; каждое из множеств *D(B), Ds(B)* включает, в частности, символ “запятая”.

Определение 2. Пусть B – произвольный к.б., $1\le i\le 10$, и множества цепочек

$$\mathbf{L(B) \subset D^+(B)\ , T^0(B), T^1(B), \ldots , T^i(B), Y^1(B), \ldots , Y^i(B) \subset Ds^+(B)}$$

являются наименьшими множествами, совместно задаваемыми правилами P[0] – P[I]. Тогда обозначим эти множества соответственно через $Lnr_i(B)$, $T^0(B)$, $Tnr_i^1(B)$, ... , $Tnr_i^i(B)$, $Ynr_i^1(B)$, ... , $Ynr_i^i(B)$ и обозначим семейство (т.е. множество), состоящее из всех этих множеств, через $Globset_i(B)$. Кроме того, пусть

$$\mathbf{T_i(B) = T_0(B) \cup Tnr_i^1(B) \cup \ldots \cup Tnr_i^i(B)} \quad , \qquad \mathbf{(3.8.1)}$$

$$Y_i(B) = Ynr_i^1(B) \cup ... \cup Ynr_i^i(B) \quad , \qquad (3.8.2)$$

$$Form_i(B) = Lnr_i(B) \cup T_i(B) \cup Y_i(B) \quad . \qquad (3.8.3)$$

Определение 3 (итоговое). Если B-произвольный к.б., то

$$Ls(B)=Lnr_{10}(B) \quad , \qquad (3.8.4)$$

$$Ts(B)=T_{10}(B) \quad , \qquad (3.8.5)$$

$$Ys(B)=Y_{10}(B) \quad , \qquad (3.8.6)$$

$$Forms(B)=Form_{10}(B) \quad . \qquad (3.8.7)$$

$$Ks(B)= (B, Rls) , \qquad (3.8.8)$$

где Rls = {P[0], P[1], ... , P[10]} .

Упорядоченная пара *Ks(B)* называется К-исчислением (концептуальным исчислением) в базисе *B*; элементы множества *Forms(B*) называются формулами, выводимыми в базисе B. Формулы из *Ls(B), Ts(B) и Ys(B)* называются соответственно *l*-формулами, *t*-формулами, *y*-формулами. Множество *l*-формул *Ls(B)* называется стандартным концептуальным языком (стандартным К-языком, СК-языком) в базисе B.

Утверждение 3.1. **Если B-произвольный к.б., то (а) множество $Lnr_0(B)$ не является пустым; (б) если $1 \le i \le 10$, то $Lnr_{i-1}(B) \subseteq Lnr_i(B)$.**

Доказательство. (а) Для любого к.б. B первичный информационный универсум X(B) включает непустое множество сортов St(B), причем $\forall s \in St(B)$, $tp(s)=\uparrow s$. Тогда при i = 0 из правила P[0] следует, что $Lnr_0(B)$ включает X(B) и, как следствие, включает St(B) . Поэтому множество $Lnr_0(B)$ непусто.

(б) Структура правил P[1] – P[10] показывает, что добавление нового правила может либо не изменить множество выводимых формул, либо его расширить. В частности, если i = 2, то в случае, когда множество функциональных символов F(B) пусто, $Lnr_{i-1}(B) = Lnr_i(B)$.

Утверждение 3.2. **Если B - произвольный к.б., то множества Ls(B), Ts(B), Ys(B) не являются пустыми.**

Доказательство

Из Утверждения 3.1 следует, что $Lnr_0(B) \subset Ls(B)$. Поэтому Ls(B) непусто. Из определения концептуально-объектной системы (к.о.с.) вытекает, что существуют такие различные переменные $v_1, v_2 \in V(B)$, что $tp(v_1) = tp(v_2) =$ [сущн]. Тогда из правила P[3] вытекает, что цепочка $(v_1 \equiv v_2)$ & P входит в

множество Ts(B), и v_1 & v_2 & ($v_1 \equiv v_2$) ∈ Ys(B), где P=P(B)-сорт "смысл сообщения" . Поэтому множества Ts(B) и Ys(B) не являются пустыми..

Утверждение 3.3. **Если B - произвольный к.б., то: (а) Если $\tau \in$ Ts(B), то τ - цепочка вида α & t, где $\alpha \in$ Ls(B), t∈ Tp(S(B)), и такое представление, зависящее от τ, единственно для каждой цепочки τ; (б) Если $\gamma \in$ Ys(B), то найдутся такое n>1 и такие цепочки $\alpha_1, \alpha_2, \ldots, \alpha_n, \beta \in$ Ls(B), что γ - цепочка вида α_1 & α_2 & ... & α_n & β; кроме того, такое представление зависящее от γ, единственно для любого γ.**

Доказательство. Справедливость этого предложения непосредственно следует из определения к.б. и из структуры правил P[0] – P[10].

Утверждение 3.4. **Пусть B - произвольный к.б., d ∈ X(B) ∪ V(B). Тогда не найдутся такие k, n , где 1≤ k ≤ 10, n>1, и такие $\alpha_1, \alpha_2, \ldots, \alpha_n \in$ Ls(B), что**

α_1 & α_2 & ... & α_n & d ∈ Ynr_{10}^{k}(B). (*)

Интерпретация. **Смысл утверждения в том, что для каждого элемента d, входящего в первичный информационный универсум X(B) или являющегося переменной из V(B), нельзя получить этот элемент d с помощью каких-либо операций, задаваемых правилами P[1] – P[10].**

Доказательство **(от противного)**

Предположим, что существуют такие к.б. B, d ∈ X(B) ∪ V(B), натуральное число k , где 1≤ k ≤ 10, n>1, $\alpha_1, \alpha_2, \ldots, \alpha_n \in$ Ls(B), что справедливо соотношение (*). Для любого такого m, что 1≤ m ≤ 10, Ynr_{10}^{m}(B) включает цепочку α'_1 & α'_2 & ... & α'_n & d', где d' не включает символ &, только в том случае, когда цепочка d' построена из цепочек $\alpha'_1, \alpha'_2, \ldots, \alpha'_n$ применением правила P[m] на последнем шаге вывода. Но тогда из структуры правил P[1] – P[10] непосредственно следует, что цепочка d' должна содержать по крайней мере два символа. Но так как d ∈ X(B) ∪ V(B), то элемент d рассматривается как символ и поэтому имеет длину 0. Мы получили противоречие из нашего предположения, что доказывает Утверждение 3.4.

Утверждение 3.5. **Пусть B - произвольный к.б., z ∈ Ls(B) \ (X(B) ∪ V(B)). Тогда существует один и только один такой набор (k, n, $y_1, y_2, \ldots, y_n$), где 1 ≤ k ≤ 10, n > 1, $y_1, y_2, \ldots, y_n \in$ Ls(B), что**

y_1 & y_2 & ... & y_n & z ∈ Ynr_{10}^k(B).

Интерпретация. **Если l-формула z не входит в (X(B) ∪ V(B)), то тогда найдутся единственное правило P[k], где 1≤ k≤ 10, и единственный такой набор l-формул y_1, y_2, ... , y_n, что цепочка z построена из "блоков" y_1, y_2, ... , y_n применением ровно один раз правила P[k].**

Справедливость Утверждения 3.5 вытекает из двух лемм, рассматриваемых ниже. Для того, чтобы сформулировать эти леммы, потребуется

Определение 4. Пусть *B*-произвольный к.б., $n \geq 1$, для $i=1,\ldots,n$ $c_i \in D(B)$, $s=c_1\ldots c_n$, $1 \leq k \leq 10$. Тогда через $lt_1(s, k)$ и $lt_2(s, k)$ обозначим количество вхождений символа '(' и символа '<', соответственно, в подцепочку $c_{1\ldots}c_k$ цепочки $s=c_{1\ldots}c_n$. Через $rt_1(s, k)$ и $rt_2(s, k)$ обозначим количество вхождений символа ')' и символа '>' в подцепочку $c_{1\ldots}c_k$ цепочки *s*. Если в подцепочку $c_1\ldots c_k$ не входит символ '(' или символ '<', то, соответственно, $lt_1(s, k) = 0$, $lt_2(s, k) = 0$, $rt_1(s, k) = 0$, $rt_2(s, k) = 0$.

Лемма 1. Пусть *B*-произвольный к.б., $y \in Ls(B)$, $n \geq 1$, для $i = 1,\ldots, n$ $c_i \in D(B)$, $y = c_1\ldots c_n$. Тогда:

(a) при $n>1$ для каждого $k = 1,\ldots, n-1$ и каждого $m = 1, 2$ $lt_m(y, k) \geq rt_m(y, k)$;

(b) $lt_m(y, k) = rt_m(y, k)$.

Лемма 2. Пусть *B*-произвольный к.б., $y \in Ls(B)$, $n > 1$, $y= c_1\ldots c_n$, где для $i=1,\ldots,n$ $c_i \in D(B)$, цепочка y включает запятую или какой-либо из символов ≡ , ∧, ∨ , и k - такое произвольное натуральное число, что $1 < k < n$. Тогда:

(а) если c_k - один из символов ≡ , ∧, ∨ , то $lt_1(y, k) > rt_1(y, k) \geq 0$;

(б) если c_k - запятая , то выполняется по крайней мере одно из соотношений $lt_1(y, k) > rt_1(y, k) \geq 0$, $lt_2(y, k) > rt_2(y, k) \geq 0$.

Доказательства Леммы 1, Леммы 2 и Утверждения 3.5. изложены в Приложении к данной книге.

Определение 5. Пусть *B* – произвольный к. б., $z \in Ls(B) \setminus (X(B) \cup V(B))$, и существует такой набор (k ,n, $y_0,\ldots, y_n$), где $1 \leq k \leq 10$, $n > 1$, и $y_1,\ldots, y_n \in Ls(B)$, что

$$y_1 \&\ldots\& y_n \& z \in Ynr_{10}^k(B).$$

Тогда упорядоченный набор вида ($k, n, y_1,\ldots, y_n$) будем называть *формообразующим набором* цепочки z.

С учетом этого определения Утверждение 3.5 говорит о том, что для произвольного к.б. B каждая цепочка $z \in Ls(B)\backslash(X(B) \cup V(B))$ имеет единственный формообразующий набор.

Утверждение 3.6. Пусть B – произвольный концептуальный базис, $z \in Ls(B)$. Тогда существует один и только один такой тип $t \in Tp(S(B))$, что $z \,\&\, t \in Ts(B)$ **Доказательство.**

Рассмотрим два возможных случая.

Случай 1.

Пусть B – произвольный к. б., $z \in X(B) \cup V(B)$, $t \in Tp(S(B))$, $tp(z) = t$. Тогда из правила P[0] следует, что $z \,\&\, t \in Ts(B)$.

Предположим, что w - такой тип из $Tp(S(B))$, что $z \,\&\, w \in Ts(B)$. Анализ правил P[0] – P[10] показывает, что такое соотношение может вытекать только из правила P[0] . Но тогда w однозначно определяется данным правилом, поэтому w совпадает с t .

Случай 2.

Пусть B – произвольный к. б., $z \in Ls(B) \setminus (X(B) \cup V(B))$. В силу Предложения 6, найдутся такие натуральные k, где $1 \le k \le 10$, $n \ge 1$, и такие y_0 , $y_{1,\ldots,}$ $y_n \in Ls(B)$, что цепочка z построена из этих элементов в результате одного применения правила P[k]. Поэтому найдется такой тип $t \in Tp(S(B))$, что $z \,\&\, t \in Tnr_{10}^{k}(B)$, и, следовательно, $z \,\&\, t \in Ts(B)$.

Из Утверждения 3.5 вытекает, что z определяет однозначным образом такие k , n , y_0 , $y_{1,\ldots,}$ y_n . Но тогда набор (k , n , y_0 , $y_{1,\ldots,}$ y_n) однозначно определяет такой тип u , что $z \,\&\, u \in Tnr_{10}^{k}(B)$. Поэтому u совпадает с t .

Таким образом, Утверждение 3.6 говорит о том, что каждой цепочке стандартного К-языка $Ls(B)$, где B - произвольный к.б., можно поставить в соответствие единственный тип t из $Tp(S(B))$.

Определение 6. Пусть B – произвольный к. б., $z \in Ls(B)$. Тогда типом l-формулы z называется такой элемент $t \in Tp(S(B))$, обозначаемый через $tpl(z)$, что $z \,\&\, t \in Ts(B)$. □

Глава 4

ИССЛЕДОВАНИЕ ВЫРАЗИТЕЛЬНЫХ ВОЗМОЖНОСТЕЙ СТАНДАРТНЫХ К-ЯЗЫКОВ

Проведем дополнительный анализ выразительных возможностей стандартных К-языков (СК-языков) по сравнению с анализом возможностей математического описания структурированных значений ЕЯ-текстов, выполненным в предыдущих параграфах**.** Набор примеров, рассмотренных выше, недостаточно полно демонстрирует реальную мощность построенной модели. Поэтому рассмотрим ряд дополнительных примеров, иллюстрирующих некоторые важные возможности СК-языков.

Если цепочка *Expr* некоторого СК-языка является семантическим представлением выражения Т на естественном языке, то такую цепочку *Expr* будем называть возможным К-представлением (КП) выражения Т.

4.1. Удобный способ описания событий

Ключевую роль в формировании предложений играют глаголы и лексические единицы, являющиеся производными от глаголов - причастия, деепричастия и

отглагольные существительные, потому что они выражают разнообразные отношения между объектами рассматриваемой предметной области.

Тематической ролью (концептуальным падежом, семантическим падежом, глубинным падежом, семантической ролью) в компьютерной лингвистике называется смысловое отношение между значением глагольной формы и значением зависящей от нее в предложении группы слов (или отдельного слова).

В таких разных языках, как русский, английский, немецкий и французский языки, можно наблюдать следующую закономерность: в предложениях с одним и тем же глаголом, обозначающим событие, явно реализуется разное количество тематических ролей, связанных со значением данного глагола. Например, пусть Т1 = “Профессор Новиков прилетел вчера” и Т2 = “Профессор Новиков прилетел вчера из Праги”. Тогда в предложении Т1 явно реализуются тематическая роль, которой можно дать название Агент1 (Агент действия), а также тематическая роль Время. При этом в предложении Т2 явно реализуются тематические роли Агент1, Время и Место1 (отношение, связывающее событие перемещения в пространстве и исходный пространственный объект).

Рассмотрим столь же гибкий способ построения СП сообщений о событиях. Для этого потребуется сделать определенное предположение о свойствах рассматриваемого концептуального базиса (к.б.) В.

Предположение 1. Множество сортов *St(B)* включает выделенный сорт *cит* (“ситуация”); множество переменных *V(B)* включает счетное подмножество V_{sit} = *{e1, e2, e3, …},* такое, что для каждого $v \in V_{sit}$ *tp(v) = cит ;* первичный информационный универсум *X(B)* включает бинарный реляционный символ *Ситуация*, такой, что *tp(Ситуация)* - цепочка {(*cит*, $\uparrow$*cит*)}. □

Смысл выражения $\uparrow$*cит* в правой части соотношения *tp(Ситуация)={(cит*, $\uparrow$*cит*)}заключается в том, что мы сможем строить выражения вида *Ситуация* (e_k *, concept),* где e_k – переменная, обозначающая конкретное событие (продажа, покупка, отлет), *concept* – простое или составное понятие, являющееся семантической характеристикой события.

Уточним, что связь между меткой ситуации и видом ситуации будет осуществляться с помощью формул вида *Ситуация (v, conc * (r_1, d_1)…(r_n, d_n)),*

где *v* – переменная типа *сит*, *conc*$\in$*X(B)*, *conc* интерпретируется как понятие, характеризующее ситуацию, $n \geq 1$, для $i=1,\ldots, n$ r_i – характеристика ситуации, d_i – значение характеристики.

Пример. Пусть *Expr1 = ∃ e1(сит) (Ситуация(e1, прилет * (Время, x1)(Агент1, нек чел * (Квалиф, профессор)(Фамилия, 'Новиков') : x2)) ∧ Раньше (x1, Сейчас))* ,

*Expr2 = ∃ e1(сит) (Ситуация(e1, прилет * (Время, x1)(Агент1, нек чел * (Квалиф, профессор)(Фамилия, 'Новиков') : x2)(Место1, нек город * (Название, 'Прага') : x3)) ∧ Раньше (x1, Сейчас))* .

Тогда легко видеть, что можно построить такой к.б. B, что для него будет справедливо Предположение 1 и выполнено соотношение *B(0, 1, 2, 3, 4, 5,7, 8, 9) => Expr1, Expr2 ∈ Ls(B), Expr1 & сообщ ∈ Ts(B), Expr2 & сообщ ∈ Ts(B)* ,

где *сообщ = P(B)* – выделенный сорт "смысл сообщения" базиса *B*.

Данный способ описания сообщений будет многократно использован в этом и последующих параграфах. При этом чаще всего, по соображениям компактности, будет опускаться квантор существования при переменной, обозначающей событие. Например, вместо формулы *Expr1* будет рассматриваться формула *Expr3* вида

*(Ситуация(e1, прилет * (Время, x1)(Агент1, нек чел * (Квалиф, профессор)*

(Фамилия, 'Новиков') : x2)) ∧ Раньше (x1, сейчас)) .

4.2. Формализация предположений о структуре семантических представлений множеств

Сообщения, вопросы, команды могут включать обозначения множеств. Для обеспечения единства подхода в разных ситуациях (при рассмотрении сообщений, команд, вопросов) к построению СП описаний множеств целесообразно ввести ряд дополнительных предположений об используемых концептуальных базисах.

В связи с тем, что обозначения множеств в текстах часто включают количественные числительные или обозначения натуральных чисел ("два

алюминиевых контейнера" и т.п.), будем полагать, что для рассматриваемого к.б. В справедливо

Предположение 2. Множество сортов *St(B)* включает выделенный сорт *нат* ("натуральное число"), первичный информационный универсум *X(B)* включает подмножество цепочек *Nt*, такое, что $Nt = \{ d_1 \ldots. d_k \mid k \geq 1$, для $i = 1,.., k$ d_i – символ из множества {'0', '1', '2', '3', '4', '5', '6', '7', '8', '9'}, и из d_1 = '0' следует, что $k = 1\}$. При этом для каждого $z \in Nt$ $tp(z) = нат$.

Потребуем также, чтобы первичный информационный универсум *X(B)* включал выделенные элементы *множ*, *Колич*, *Кач-состав*, *Предм-состав*, интерпретируемые следующим образом: *множ* – это обозначение понятия "конечное множество", *Колич* – имя одноместной функции "Количество элементов множества", *Кач-состав* – имя бинарного отношения "Качественный состав множества", *Предм-состав* – имя бинарного отношения "Предметный состав множества".

Предположение 3. Первичный информационный универсум *X(B)* включает элементы *множ, Колич, Кач-состав, Предм-состав,* такие, что *tp(множ) = ↑{[сущн]}, tp(Колич) = {({[сущн]},нат)} , tp(Кач-состав) = {({[сущн]}, [пон])}, tp(Предм-состав) = {({[сущн]}, [сущн])}* , где *[сущн], [об], [пон]* – базовые типы “сущность” , “объект”, “понятие” (см. параграф 1.6).

Рассмотрим назначение выделенных элементов универсума *X(B)*, упоминаемых в Предположении 3.

Используя элементы *множ, Колич* и произвольную цепочку *numb* из *Nt*, мы сможем построить СП выражения "некоторое множество, содержащее *numb* элементов" в виде *нек множ * (Колич, numb)*, где *нек = ref(B)* – квантор референтности рассматриваемого концептуального базиса *B*.

Назначение бинарного реляционного символа *Кач-состав* заключается в следующем. Пусть *v* – переменная, обозначающая некоторое множество, и *conc* – простое или составное обозначение понятия. Тогда выражение *Кач-состав (v, conc)* обозначает высказывание "Каждый элемент множества *v* квалифицируется понятием *conc*", и это высказывание может быть истинным или ложным. Примерами выражений этого вида являются *Кач-состав(S1, контейнер1)*,

*Кач-состав(S2, статья1), Кач-состав(S3, контейнер1 * (Материал, алюминий)), Кач-состав(S4, статья1 * (Область1, биология)).*

С другой стороны, символ *Кач-состав* будет применяться и при построении составных обозначений множеств вида *ref множ * (Кач-состав, conc) : v,*
где *ref* – квантор референтности, *conc* – простое или составное обозначение понятия, *v* – переменная. В частности, концептуальный базис *B* можно выбрать так, чтобы язык *Ls(B)* включал выражения

*нек множ * (Кач-состав, контейнер1) : S1, нек множ * (Кач-состав, статья1) : S2, нек множ * (Кач-состав, контейнер1 * (Материал, алюминий)) : S3, нек множ * (Кач-состав, статья1 * (Область1, биология)) : S4.*

Фрагмент текста, обозначающий множество, может представлять собою явное перечисление элементов множества. Таким, в частности, является текст "Два заказчика, АО "Радуга" и ТОО "Зенит", не оплатили сентябрьские поставки".

Бинарный реляционный символ *Предм-состав* предназначен, в частности, для построения выражений вида *Предм-состав (v, ($x_1 \wedge x_2 \wedge \ldots \wedge x_n$))*,
где *v*, x_1 , x_2 ,…, x_n – переменные, причем *v* обозначает множество, а x_1 ,…, x_n – это обозначения всех элементов, входящих в состав множества *v*.

Например, будем считать, что выражение *(Предм-состав (y1, ($x_1 \wedge x_2$)) ∧ Явл (x1, АО) ∧ Явл (x2, ТОО) ∧ Имя (x1, "Радуга") ∧ Имя (x2, "Зенит"))* является СП высказывания "Множество y1 состоит из АО "Радуга" и ТОО "Зенит", причем первая организация обозначается через x1, а вторая – через x2.

В то же время мы должны иметь возможность (если это необходимо) строить формулы вида *ref множ * (Предм-состав, ($x_1 \wedge x_2 \wedge \ldots \wedge x_n$)) : y1,* где *ref* – квантор референтности, и *y1, x1, x2* – переменные типа *[сущн]* (базовый тип "сущность"). Например, мы должны располагать возможностью построения выражения *нек множ * (Предм-состав, (x1 ∧ x2)) : y1* .

Пример. Рассмотрим выражения Вр1 = "3 контейнера с керамикой из Индии" и Вр2 = "Партия керамики, состоящая из коробок с номерами 3217, 3218, 3219". Тогда можно построить такой к.б. *B,* для которого будут выполнены Предположения 2, 3, и *Ls(B)* включает формулы
*(1) нек множ. * (Колич, 3) (Кач-состав, Контейнер 1 * (Содерж1, нек множ **

*(Кач-состав, изделие * (Вид, керамика) (Страна, Индия)))),*

*(2) (нек партия2 * (Колич, 3)(Предм-состав, (нек коробка1 * (Номер, 3217) : x1 ∧ нек коробка1 * (Номер, 3218) : x2 ∧ нек коробка1 * (Номер, 3219) :x3)) :S1 .*

Построенные формулы будем интерпретировать как возможные КП выражений Вр1 и Вр2; здесь *x1, x2, x3* – метки коробок, *S1* – метка партии.

4.3. Построение семантических представлений вопросов с ролевыми вопросительными словами

Среди всех вопросительных местоимений и наречий можно выделить подмножество, включающее, в частности, слова “кто”, “что”, “кому”, “чем”, “когда”, “откуда”. Чтобы сформулировать свойство, выполняющееся для каждого элемента этого подмножества, введем обозначение *nil* для пустого предлога. Если в каком-либо вопросе некоторое вопросительное местоимение *qswd* употреблено без предлога, то условимся говорить, что этому местоимению *qswd* в данном предложении соответствует пустой предлог.

Для каждого местоимения qswd из рассматриваемого подмножества найдется предлог prep (возможно, он не является единственным), что паре (prep, qswd) соответствует некоторая тематическая роль *role*. Например, парам *(nil, кому), (для, кого), (от, кого),* могут соответствовать тематические роли *Адресат, Адресат, Источник1*. Таким образом, разным парам *(nil, кому), (для, кого)* соот-ветстветствует одна тематическая роль *Адресат*, поскольку в равной степени пра-вильными являются фразы “Кому прислана книга?” и “Для кого прислана книга ?”.

Местоимения и наречия, входящие в указанное подмножество, будем называть *ролевыми вопросительными словами.*

Предположение 4. Первичный информационный универсум *X(B)* концептуального базиса *B* включает символ *Вопрос*, и *tp(Вопрос) = {([сущн], P)}* , где *tp = tp(B)* - отображение, задающее тип информационной единицы, *[сущн]* – базовый тип “сущность”, *P = P(B)* – выделенный сорт “смысл сообщения”.

Пусть для к.б. *B* выполнено Предположение 4. Тогда СП вопроса с *n* ролевыми вопросительными словами можно представить в виде *Вопрос* (v_1 , *A*) при *n = 1* и в виде *Вопрос* (($v_1 \wedge .... \wedge v_n$), *A*) при *n > 1* , где *A* - формула, зависящая от переменных v_1 , ... , v_n и отображающая содержание высказывания (т.е. являющаяся семантическим представлением высказывания).

Пример 1. Пусть B1 = “Откуда поступил трехтонный алюминиевый контейнер?”, *Expr1* – цепочка вида *Вопрос (x1 , (Ситуация (e1, поступление2 * (Объект1, нек контейнер * (Вес, 3/тонна)(Материал, алюминий) : x2)(Место1, x1)(Время, t1)) ∧ Раньше(t1, сейчас)))* . Тогда нетрудно построить такой к.б. *B*, что для *B* выполняются Предположение 1 и Предположение 4, *P(B)* = сообщ, и *B (0, 1, 2, 3, 4, 5, 7, 8) => Expr1 ∈ Ls(B) , Expr1& сообщ ∈ Ts(B)* .

Цепочка *Expr1* является возможным КП вопроса B1. В этой цепочке символы *x1, x2, e1, t1* являются переменными, *поступление2* - информационная единица (другими словами, семантическая единица), соответствующая существительному “поступление” и передающая значение “перемещение некоторого физического объекта на пространственный объект” (в отличие от значения “поступление абитуриента в учебное заведение”).

Пример 2. Пусть B2 = “Откуда и когда поступил трехтонный алюминиевый контейнер?”. Тогда КП вопроса B2 может являться выражением *Вопрос ((x1 ∧ t1) , (Явл1 (e1, поступление2 * (Объект1, нек контейнер * (Вес, 3/тонна)(Материал, алюминий) : x2)(Место1, x1)(Время, t1)) ∧ Раньше(t1, Сейчас)))* .

4.4. Семантические представления вопросов о количестве предметов и о количестве событий

Предположение 5. Первичный информационный универсум *X(B)* вида *(S, Ct, Ql),* где *Ql* – система кванторов и логических связок вида *(ref,* int_1*,* int_2*, eq, neg, binlog, ext)* , включает элементы *произв, все, Элем*, такие, что *tp(произв) =* int_1*,* *tp(все) =* int_2*, tp(Элем) = {([сущн], {[сущн]})}* .

Ээлементы *произв, все, Элем* интерпретируются как информационные единицы “произвольный” (“каждый”), “все” и “Элемент множества” (имя отношения “Быть элементом множества”).

Следует заметить, что int_1 и int_2 — это выделенные элементы множества сортов *St(B).* По определению (см. параграф 2.8), элементы int_1 и int_2 являются типами интенсиональных кванторов соответственно первого и второго видов.

Пример 1. Пусть В1= «Сколько экземпляров книг А.П.Сомова имеется в библиотеке?». Тогда можно определить такой к.б. *B*, что для *B* выполняются Предположение 4 и Предположение 5, и цепочка

Вопрос(x1, (x1≡ Колич (все экземпляр1(Информ-объект, произв книга * (Автор, нек чел* (Инициалы, ‘А.П.’)(Фамилия, ‘Сомов’) : x2) : x3) (Место-хранения, нек библиотека: x4))))*

входит в *Ls(B).* Поэтому данное выражение является возможным К-представлением вопроса В1.

Пример 2. Если В2 = «Сколько человек участвовало в создании статистического сборника?», то возможным К- представлением В2 является выражение

Вопрос(x1, ((x1≡ Колич(все чел(Элем, S1))) ∧ Описание(произв чел* (Элем, s1) : y1, (Ситуация(e1, участие1* (Агент1, y1)(Время, x2) (Вид-деятельности, создание1* (Продукт1, нек сборник1* (Область1, статистика)))) ∧ Раньше(x2, #Сейчас#))))) .*

Пример 3. Пусть В3 = «Сколько книг поступило в январе этого года в библиотеку № 18?». Тогда возможным К-представлением В3 является формула

*Вопрос(x1, ((x1≡ Колич (все книга * (Элем, S1))) ∧ Описание (произв книга* (Элем, S1) : y1, (Ситуация(e1, поступление2* (Объект1, y1)(Время, <01, текущий- год>)(Место2, нек библиотека * (Номер, 18) : x2)))) .*

Пример 4. Вопрос В1= «Сколько раз Иван Михайлович Семёнов летал в Мексику?» может иметь следующее возможное КП:

*Вопрос(x1, (x1≡ Колич (все полёт *(Агент1, нек чел * (Имя, ‘Иван’) (Отчество, ‘Михайлович’)(Фамилия, ‘Семёнов’): x2)(Место2, нек страна* (Название, ‘Мексика’):x3)(Время, произв момент *(Раньше, #сейчас#))))) .*

4.5. Семантические представления вопросов с формами вопросительно-относительного местоимения “какой”

Метод, предложенный выше для построения К-представлений вопросов с ролевыми вопросительными словами, можно использовать и для построения КП вопросов с различными формами местоимения “какой”.

Пример 1. Пусть В1= «Какое издательство опубликовало роман «Ветры Африки»?». Тогда КП вопроса В1 может являться цепочкой

*Вопрос(x1, (Ситуация(e1, опубликование * (Время, x2) (Агент2, нек издательство: x1) (Объект3, нек роман1* (Название, ‘Ветры Африки’) :x3)) ∧ Раньше(x2, #сейчас#)))* .

Пример 2. Пусть В2= «С какими зарубежными издательствами сотрудничает писатель Игорь Сомов?». Тогда КП В2 может являться формулой

*Вопрос (S1, (Кач-состав (S1, издательство * (Вид-географич, зарубежное))*

∧ Описание(произв издательство (Элем, S1) : y1, Ситуация(e1, сотрудничество * (Агент1, нек чел* (Профессия, писатель)(Имя, ‘Игорь’)(Фамилия, ‘Сомов’): x1)(Организация1, y1)(Время, #сейчас#)))))* .

4.6. Построение семантических представлений вопросов общеудостоверительного актуально-синтаксического типа

Вопросами общеудостоверительного актуально-синтаксического типа в лингвистике называются вопросы с ответом “Да“ или “Нет“. Такие вопросы задаются для того, чтобы в целом удостовериться в правильности имеющейся у спрашивающего информации (Воробьева, Панюшева, Толстой 1975).

Оказывается, что предложенную выше форму отображения смысла вопросов с ролевыми вопросительными словами можно использовать и для построения СП общих вопросов. Для этого каждый такой вопрос будем интерпретировать как просьбу указать истинное значение некоторого высказывания. Например, вопрос В1 = "Является ли Гент городом Бельгии" можно интерпретировать как просьбу найти истинностное значение высказывания "Гент является одним из городов Бельгии". Для реализации этой идеи введем

Предположение 6. *St(B)* включает выделенный сорт *лог*, называемый "логическая величина"; *X(B)* включает различные элементы *ист*, *ложь*, причем *tp(ист) = tp(ложь) = лог* , *F(B)* включает одноместный функциональный символ *Ист-знач*, такой, что *tp(Ист-знач) = {(P, лог)}* , где *P = P(B)* – выделенный сорт "смысл сообщения".

Пример 1. Если для рассматриваемого концептуального базиса B выполняется указанное предположение, то КП вопроса B1 = "Является ли Гент городом Бельгии?" может являться формулой

*Вопрос(x1, (x1 ≡ Ист-знач (Элем(нек город * (Назв, "Гент") : x2,*

*Города(нек страна * (Назв, "Бельгия") : x3))))).*

В этой формуле символ *Города* интерпретируется как имя одноместной функции, ставящий в соответствие стране множество всех городов этой страны, *нек* – квантор референтности *ref(B); x1, x2, x3∈ V(B).*

Пример 2. Пусть B2= «Проходила ли в Азии международная научная конференция «COLING»?». Тогда К-представлением вопроса B2 может являться формула

Вопрос(x1, (x1≡ Ист-знач ((Ситуация (e1, прохождение2(Событие, нек конф* (Вид1, междун) (Вид2, научн) (Название, 'COLING') : x2) (Место, нек континент* (Название, 'Азия') :x3) (Время, x4)) ∧ Раньше(x4, #сейчас#)))) .*

4.7. Отображение смысловой структуры команд

Будем использовать две основные идеи. Во-первых, когда мы говорим о команде (или о приказе, распоряжении и т.д.), то всегда подразумеваем, что имеется одна интеллектуальная система, формирующая команду (обозначается выражением *#Оператор#),* и другая интеллектуальная система (или же конечное множество интеллектуальных систем), которая должна выполнить команду (обозначается выражением *#Исполнитель#).* Во-вторых, глагол в повелительном наклонении или неопределенную форму глагола будем заменять соответствующим отглагольным существительным.

Предположение 7. Множество сортов *St(B)* включает выделенные элементы *интс* (сорт "интеллектуальная система"), *мом* (сорт "момент времени");

первичный информационный универсум *X(B)* включает элементы *Команда, #Оператор#, #Исполнитель#, #сейчас#,* такие что *tp(Команда*) = {(*интс, интс, мом, ↑сит*)}, *tp(#Оператор#) = tp(#Исполнитель#) = интс, tp(#сейчас#) = мом*.

Пример. Пусть К1 = "Доставь ящик с деталями на склад № 3", где К1 - команда, отданная оператором гибкой производственной системы интеллектуальному транспортному роботу. Тогда базис В можно определить так, чтобы выполнялись Предположение 1, Предположение 7 и соотношение

B(0, 1, 2, 3, 4, 5, 8)=> Команда(#Оператор#, #Исполнитель#, #сейчас#, доставка1(Объект1, нек ящик * (Содерж1, нек множ * (Кач-состав, деталь)) : x1)(Место2, нек склад * (Номер, 3) : x2)) ∈ Ls(B).*

4.8. Представление теоретико-множественных отношений и операций на множествах

Пример 1. Пусть Т1а = "Намюр – один из городов Бельгии". Тогда рассмотрим текст Т1б = "Намюр входит в множество всех городов Бельгии".

Пусть *E1 = Элем (нек город * (Назв, 'Намюр') : x1, Города(нек страна * (Назв, 'Бельгия') : x2))* . Тогда ∃ такой к.б. *B1*, что *B1(0, 1, 2, 3, 8, 1, 5, 0, 1, 2, 3, 8, 1, 5, 2, 4) => E1 ∈ Ls(B1).* При построении нужно предполагать, что *Элем, страна, город, нек ∈ X(B), tp(Элем) = {([сущн], {[сущн]})}, tp(страна) = tp(город)= ↑ простр.об; Города ∈ F1(B1), tp(Города) = {(простр. об, {простр. об})}, нек = ref(B)* – квантор референтности базиса *B.*

Пример 2. Пусть Т2а = "Включи контейнер № 4318 в партию, отправляемую в Тамбов". Преобразуем Т2а в Т2б ="Некоторый оператор распорядился включить контейнер № 4318 в некоторую партию, отправляемую в город Тамбов". Тогда построим КП текстов Т2а и Т2б в виде

Команда((#Оператор#, #Исполнитель#, #сейчас#, включение1 (Объект1, нек контейнер * (Номер, 4318) : x1) (Целевое.множество, нек партия2 * (Место-назн, нек город*(Название, 'Тамбов') : x2) : S1)),*

где *S1* - метка партии продукции. Аналогично можно представить распоряжения о разделении множества объектов на несколько частей и об объединении

нескольких множеств в одно, например, при перегрузке деталей из нескольких ящиков в один.

4.9. Представление смысла фраз с придаточными предложениями цели и с косвенной речью

Пример 1. Пусть T1 = "Сергей поступил в МИЭМ, чтобы получить специальность "Прикладная математика" ", и *Sr1 = (Ситуация (e1, поступление1* (Агент, нек чел* (Имя, "Сергей") :x1)(Уч.заведение, нек вуз * (Название, 'МИЭМ') : x2)(Время, t1) (Цель, получение1* (Квалификация, нек специальность*(Название, 'прикладная математика') : x3)) ∧ Раньше(t1, #сейчас#)))).* Тогда ∃ такой.к.б. *B*, что *B(0, 1, 2, 3, 4, 5, 7, 8) => Sr1 ∈ Ls(B), Sr1 & Сообщ ∈ Ts(B).*

Пример 2. Пусть T2 = "Директор сказал, что на февраль запланирована реорганизация фирмы" и

*Sr2 = (Ситуация (e1, устное-сообщение * (Агент1, Директор(нек организация: x1)) (Время, t1)(Содержание1, Планируется(нек реорганизация* (Объект2, нек фирма: x1), Ближайший-месяц(февраль, t1)))) ∧ Раньше(t1, #сейчас#)) .*

Тогда легко построить такой к.б. *B*, что выполняются соотношения

B(0,2) => Ближайший-месяц(февраль, t1) ∈ Ls(B), Ближайший-месяц(февраль, t1) & врем.интервал ∈ Ts(B) ; B(0,1,2,4,5,7, 8) => Sr2∈ Ls(B), Sr2 & сообщ ∈ Ts(B).

4.10. Явное представление причинно-следственных отношений, передавамых дискурсами

Как уже отмечалось выше, в компьютерной и теоретической лингвистике дискурсом, или связным текстом, называется последовательность взаимосвязанных по смыслу предложений (полных или неполных). Соответствие между группами слов из текста и теми объектами, событиями, процессами, смыслами, которые эти группы слов обозначают, называется референтной структурой текста. СК-языки предоставляют широкие

возможности описания смысловой структуры дискурсов, в том числе их референтной структуры.

Пример. Пусть Т1 = "Первокурсник Петр Сомов не заметил, что расписание изменилось, поэтому он пропустил первую лекцию по линейной алгебре". В этом тексте проявляется, в частности, следующая особенность дискурсов: личное местоимение "он" используется вместо более длинного сочетания "первокурсник Петр Сомов". Говорят, что у этого последнего выражения и местоимения "он" есть один и тот же референт - некоторый человек, студент вуза.

Чтобы явно указать референтную структуру текста, нужно связать метки с сущностями, обозначаемыми некоторыми группами слов из этого текста или неявно упоминаемыми в тексте. Сделаем это таким образом: неявно упоминаемое учебное заведение – метка *x1*; “Первокурсник Петр Сомов”, “он” - метка *x2*; “Расписание” - метка *x3*; “первую лекцию по линейной алгебре” - метка *x4*; “не заметил” - метка *e1* (событие); “изменилось” - метка *e2*; *e3* – метка ситуации, описываемой первым предложением из Т1; “пропустил” - метка *e4* (событие). Будем полагать, что СП текста Т1 должно включать фрагмент *Причина(e3, e4).* Пусть

*Sr1 = ((Ситуация (e1, ¬ обращение-внимания * (Агент1, нек чел*(Имя, 'Петр')(Фам, ‘Сомов’)(Квалиф, студент*(Курс,1)(Уч-заведение, x1)) : x2)(Время, t1)(Объект-внимания, e2)) ∧ Раньше(t1, #сейчас#) ∧ Ситуация (e2, изменение * (Предмет, нек расписание : x3)(Время, t2)) ∧ Раньше(t2, t1)) : P1 ∧ Характеризует(P1, e3)).*

Тогда *Sr1* - возможное КП первого предложения П1 дискурса Т1.

Пусть *Sr2 = ((Ситуация (e4, пропуск1 * (Агент1, x2)(Объект3, нек лекция * (Дисциплина, лин-алгебра)(Уч-заведение, x1) : x4)(Время, t3)) ∧ Раньше(t2, #сейчас#).* Тогда *Sr2* - возможное КП второго предложения П2 из дискурса Т1.

. Пусть *Srd1 = (Sr1∧ Sr2 ∧ Причина(e3, e4)).* Тогда *Srd1* - возможное СП дискурса Т1, являющееся К-представлением текста Т1.

4.11. Построение семантических представлений дискурсов со ссылками на смысл фраз и более крупных частей текста

Пример. Пусть Т1 = "АО "Радуга" подпишет контракт до 15 декабря. Об этом сообщил заместитель директора Игорь Панов". Здесь сочетание "об этом" обозначает ссылку на смысл первого предложения дискурса Т1. Пусть

*Sr1 = (Ситуация (e1, подписание1 * (Агент1, нек организация * (Тип, АО) (Название, "Радуга"): x1)(Время, t1)(Объект3, нек контракт1: x2)) ∧ Раньше(t1, 15/декабрь/текущий-год))* ,

Srd2 = (Sr1 : P1 ∧ Ситуация (e2, сообщение1 (Агент1, нек чел * (Имя, 'Игорь')(Фамилия, 'Панов') : x3)(Время, t2)(Содержание2, P1)) ∧ Раньше (t2,, #сейчас#) ∧ Зам.директора(x3, нек орг : x4)))*.

Тогда найдется такой к.б. *B*, что *B(0,1, 2, 3, 4, 5, 7, 8) => Srd2 ∈ Ls(B), Srd2 & сообщ ∈ Ts(B)*.

Правило P[5] позволяет приписать переменную *v* к СП *Sr* произвольного повест-вовательного текста и получить формулу *Sr : v* , где *v* - произвольная переменная сорта *P(B)* - сорта "смысл сообщения". Поэтому выражениям "об этом", "этот метод", "этот вопрос" и т.д. будет соответствовать переменная *v* сорта *P(B)* в СП всего дискурса (так же, как и в последнем примере).

4.12. Представление фрагментов знаний о мире

Пример 1. Пусть Т1 = "Понятие "молекула" используется в физике, химии, биологии." Можно определить такой к.б. *B*, что множество сортов *St(B)* включает элемент *область1* и первичный информационный универсум *X(B)* включает элементы *область1, цепочка, понятие, "молекула", Использ, Имя-понятия, физика, нек., химия, биология,* причем типы этих элементов задаются соотношениями

tp(понятие) = [↑пон], tp("молекула") = цепочка, tp(физика) = tp(химия) = tp(биология) = область1, tp(Использ) = {([пон], область1)}, tp(Имя-понятия) = {([пон], цепочка)}.

Пусть *нек* — квантор референтности базиса *B*, *Использ* и *Имя-понятия* — бинарные реляционные символы, не являющиеся именами функций, и

s_1 = *Имя-понятия (нек понятие, “молекула”),*

s_2 = *понятие * (Имя-понятия, “молекула”),*

s_3 = *Использ (нек понятие * (Имя-понятия, “молекула”),*

(физика ∧ химия∧ биология)).

Тогда *B(0, 1, 4) => s_1∈Ls(B); B(0, 1, 4, 8) => s_2∈Ls(B); B(0, 1, 4, 8, 1, 0, 7, 4) => s_3∈Ls(B).* Построенная формула s_3 является возможным КП для определения *T1*.

Пример 2. Пусть T2 = “Тинейджер — это человек в возрасте от 12 до 19 лет”; s –цепочка *((тинейджер ≡ человек * (Возраст, x1)) ∧ ¬ Меньше(x1, 12/год) ∧ ¬ Больше(x1, 19/год)).* Тогда s — возможное КП для T2.

Пример 3. В работе (Nebel, Peltason 1991) сформулировано определение: “Малое и среднее предприятие (*sme*) — это компания с числом служащих не более 50”. Это определение может иметь, в частности, следующие К-представления:

Определение(sme, ∀x1(компания1)(Явл1(x1, sme) ≡

¬Больше(Колич (Персонал (x1)), 50))) ,

((sme ≡ компания1(Описание, P1))∧(P1 ≡ ∀x1(компания1)*

(Явл1(x1, sme) ≡ ¬Больше(Колич(Персонал (x1)), 50)))) .

4.13. Объектно-ориентированные представления фрагментов знаний

Используя стандартные К-языки, мы можем строить сложные описания объектов и множеств объектов. Например, мы можем построить следующее К-представление описания международного журнала “Informatica”:

нек межд-науч-журнал (Название, 'Informatica’)(Страна, Словения)*

(Город, Любляна) (Области, (иск-интеллект ∧ когнитивная-наука

∧ базы-данных)) : k225 ,

где *k225* — метка модуля знаний с данными об этом журнале.

Постановка задачи, изложенная в параграфе 3.1, предусматривает возможность строить с помощью новых формальных средств концептуальные

представления текстов как информационные объекты, отражающие не только смысл, но и значения внешних характеристик текста (метаданные): авторов, дату, области применения изложенных результатов и т. д.

Пример. Используя идею построения К-представлений разнообразных объектов, проиллюстрированную на примере модуля знаний с данными о журнале "Informatica", мы можем построить модуль знаний, содержащий теорему Пифагора и указывающий ее автора и предметную область. Например, подобный модуль может быть следующим выражением некоторого СК-языка:

нек информ-объект (Вид, теорема)(Область, геометрия)(Автор, Пифагор)*
(Содержание, ∀x_1(геом) ∀x_2(геом) ∀x_3(геом) ∀x_4(геом)Если-то((Явл(x_1, прямоугольн) ∧ Гипотенуза(x_1, x_2)∧ Катет((x_3∧x_4), x_1)),
(Квадрат(Длина(x_2)) ≡Сумма(Квадрат(Длина(x_3)), Квадрат(Длина(x_4)))))) : k81.

4.14. Сравнение выразительных возможностей СК-языков с возможностями основных известных подходов к формальному представлению содержания ЕЯ-текстов

4.14.1. Сравнение с основными подходами, разработанными в нашей стране

Основными средствами формального представления содержания ЕЯ-текстов, разработанными в нашей стране и использовавшимимя в 1990-е – 2000-е годы для проектирования лингвистических процессоров, являются, помимо предложенных автором данной дисертации стандартных К-языков (СК-языков), расширенные семантические сети (Кузнецов 1976 - 1989; Кузнецов и др. 2000; Кузнецов, Мацкевич 2001, 2003; Кузнецов, Шарнин 2003; Kuznetsov, Matskevich 2002; Соловьева, Сомин 1993), формальные выражения, предоставляемые компьютерной семантикой русского языка, и неоднородные семантическте сети (Осипов 1990, 1997).

Расширенную семантическую сеть (РСС) можно представить как конечное множество выражений вида $R(c_1, c_2, \ldots, c_n, d)$ **,** где $n \geq 1$, R – имя n-арного отноше-ния, $c_1, c_2, \ldots, c_n$ - атрибуты отношения R , d – метка, являюшаяся уникальным именем (в рамках рассматриваемой базы данных) выражения $R(c_1,$

c_2, …, c_n) . Выражения вида $R(c_1, c_2, \ldots, c_n, d)$ называются элементарными фрагментами (ЭФ).

Каждый ЭФ вида $R(c_1, c_2, \ldots, c_n, d)$ можно аппроксимировать выражением некоторого СК-языка $R(c_1, c_2, \ldots, c_n) : d$, где при построении формулы $R(c_1, c_2, \ldots, c_n)$ на последнем шаге применялось правило P[4] , а формула $R(c_1, c_2, \ldots, c_n) : d$ построена в результате применения правила P[5] к операндам $R(c_1, c_2, \ldots, c_n)$ и d , причем d является переменной.

Использование элементарных фрагментов обеспечивает большую однородность представления информации в виде РСС, что создает предпосылки для унификации процедур обработки знаний, представленных с помощью РСС. Однако принципиальным недостатком использования РСС для отображения содержания текстов является огромный разрыв между структурой ЕЯ-текстов и структурой их семантических представлений. В связи с этим, располагая только аппаратом РСС, разработчик семантико-синтаксического анализатора ЕЯ-текстов остается один на один с многочисленными проблемами алгоритмизации перехода от ЕЯ-текста к его семантическому представлению.

Исходный запас идей для развития компьютерной семантики русского языка (КСРЯ) был изложен в 5-й главе монографии (Тузов 1984). В последующие годы эти идеи получили развитие, в частности, в публикациях (Тузов 2001; Каневский, Тузов 2002; Лезин , Тузов 2003).

Центральная идея КСРЯ заключается в следующем. Каждое слово русского языка (РЯ) интерпретируется как название (имя) функции, связанной с этим словом и называемой его семантикой. Семантическое представление предложения является суперпозицией функций, причем в качестве аргументов исходных функций берутся обозначения понятий (называемые лексемами).

Анализ публикаций по КСРЯ позволяет заметить, что более естественно было бы говорить в таких ситуациях о суперпозиции функций и отношений. Например, выражения *Caus(x, y)* (x является причиной события y), *Loc(x, y)* (x расположен в y) наиболее естественно рассматривать как атомарные формулы, в которых элементы *Caus*, *Loc* интерпретируются либо как имена бинарных отношений, либо как имена бинарных предикатов.

Выражения семантического языка КСРЯ можно аппроксимировать выражениями СК-языков, полученными применением правил Р[2] и Р[4], предназначенных для использования имен функций и имен n-арных отношений ($n \geq 1$), к исходным цепочкам вида *ref concept* , где *ref* – квантор референтности (информационная единица, соответствующая слову “некоторый”), *concept* – обозначение понятия из первичного информационного универсума.

Например, в работе (Тузов 2001) семантическое представление предложения “Собака охраняет кофейную плантацию” является выражением *Oper09_a1(СОБАКА$14224112~!%1,ОХРАНА$182036(Rel_o1(ПЛАНТАЦИЯ$12411~!%1,КОФЕ$14/112))).*

К-представлением этого предложения может быть, например, выражение *∃ e1(ситуация) Является (e1, охрана1 * (Агент1, нек собака)(Объект1, нек плантация * (Растения, нек множ * (Кач-состав, дерево * (Вид, кофейн))))))* .

СК-языки удобны и для построения шаблона семантической модели предложения (см. Лезин, Тузов 2003). Например, шаблон семантической модели предложения “Вручая книгу, старик окинул мальчика быстрым оценивающим взглядом” является последовательностью формальных выражений, включающей выражения *КНИГА $14110 : X1 , СТАРИК $12411 : X2, ВРУЧЕНИЕ $15210 : X3 (Oper, СУБЪЕКТ.X2, ОБЪЕКТ.X1, АДРЕСАТ.Z3).*

Эти выражения можно аппроксимировать К-цепочками *Явл(X1 , книга), Явл (X2, старик) , Явл (e1 , вручение1 * (Агент, X2)(Объект1, X1)(Адресат, Z3)).*

Аппарат СК-языков, предложенный в данной книге, обладает следующими основными преимуществами по сравнению с КСРЯ:

(а) ориентирован на построение СП не только отдельных предложений, но и связных текстов (в частности, позволяет отображать ссылки на смысл фраз и более крупных фрагментов дискурса); (б) позволяет строить формальные аналоги сложных составных обозначений понятий и множеств; (в) предоставляет возможность отображения смысловой структуры предложений со словом “понятие”, что необходимо для формального представления информации энциклопедического характера; (г) позволяет строить формальные представления составных целей; (д) конструктивно отражает существование

нескольких дополнительных способов использования логических связок в русском, английском и многих других языках по сравнению с языком логики предикатов.

Этими же основными преимуществами аппарат СК-языков обладает и по сравнению с аппаратом неоднородных семантических сетей (Осипов 1990, 197).

Таким образом, аппарат СК-языков значительно расширяет возможности формального отображения содержания ЕЯ-текстов по сравнению с КСРЯ и по сравнению с аппаратом неоднородных семантических сетей.

Обсуждавшиеся выше аппарат расширенных семантических сетей, компьютерная семантика русского языка и аппарат неоднородных семантических сетей. нашли применение в нескольких проектах разработки ЛП.

Совсем недавно, в последние три года, в публикациях С.В. Елкина и С.С. Елкина был предложен принципиально иной подход к построению семантических языков, истоки которого лежат не в области проектирования ЛП, а в философии. Этот подход рассматривает проблему формализации семантики понятий. Для решения этой проблемы предложен т.н. открытый семантический язык SL (Елкин 2003), построенный на основе информационного исчисления (Елкин С.В, Елкин С.С 2002а, 2002б).

Если не анализировать математическую сущность данного подхода к моделированию семантики понятий, то некоторые суждения из перечисленных непосредственно выше работ могут создать впечатление, что открытый семантический язык SL может широко использоваться для отображения смысла ЕЯ-текстов в компьютерных интеллектуальных системах. Например, в работе (Елкин С.В, Елкин С.С 2002б) отмечается, во-первых, что “наиболее полно проблема понимания текста на естественном языке может быть разрешена при помощи семантического языка” (с. 97). Во-вторых, что ряд особенностей универсального сетевого языка UNL (см. параграф 1.1) требуют его доработки или принципиальных изменений, эти особенности явились источником идей для создания языка SL.

Поскольку язык UNL разрабатывался в качестве языка-посредника для устранения языкового барьера между пользователями сети Интернет из разных стран мира, можно, на первый взгляд, сделать умозаключение о больших

выразительных возможностях открытого семантического языка SL с точки зрения построения семантических представлений предложений и дискурсов на ЕЯ. Однако внимательный анализ определения семантического языка SL показывает, что такой вывод был бы ошибочным: в действительности выразительные возможности семантического языка SL являются весьма ограниченными.

Представляется, что глобальной причиной этого является отсутствие даже в постановке задачи исследования закономерностей организации поверхностной и смысловой структуры предложений и дискурсов из широко распространенных естественных языков. Уровень рассмотрения проблематики является философским. Например, в работе (Елкин С.В 2003) отмечается, что “внешние отношения ПРИТЯЖЕНИЯ и ОТТАЛКИВАНИЯ связаны с внутренними процессами – ЛЮБОВЬЮ и НЕНАВИСТЬЮ. Будем описывать чувства и эмоции следующим образом: ($С_1$ * О * $С_2$) = Эм , (С * $О_2$* $О_1$) = Эм , ($С_1$ * $С_2$ * $С_3$) = Эм “.

Как следствие, данный подход не может рассматриваться в качестве эффективного теоретического инструмента для проектирования семантико-синтаксических анализаторов.

4.14.2. Сравнение с основными зарубежными подходами

В последнее десятилетие у зарубежных исследователей, работающих в области компьютерной лингвистики, по сравнению с 1980-ми годами значительно усилился интерес к методам формального исследования семантики ЕЯ-тестов. В этот период наибольшей популярностью пользовались три подхода: теория представления дискурсов (ТПД), возникшая в начале 1980-х годов (Kamp 1981), теория концептуальных графов (ТКГ), своим появлением обязанная работам (Sowa 1984, 1991), и эпизодическая логика (ЭЛ), предложенная Л. Шубертом и Ч.Х. Хуан (Schubert 1999, 2000; Shubert, Hwang 1989, 2000; Hwang 1992; Hwang, Schubert 1993a – 1993b).

Анализ показывает, что структура любых формальных выражений, использовавшихся в теории концептуальных графов или эпизодической логике

для отображения содержания ЕЯ-текстов и представления знаний о мире, может быть аппроксимирована выражениями стандартных К-языков (СК-языков). Например, выражение *[книга : {*} @ 50]* , являющееся в ТКГ СП словосочетания “50 книг“, может быть аппроксимировано К-формулой *нек множ * (Колич-элем, 50)(Качеств-состав, книга)* , где *нек* – информационная единица (квантор референтности), соответствующая словам с лексемой “некоторый”, *Колич-элем* и *Качеств-состав* – бинарные реляционные символы, обозначающие отношения “Количество элементов множества” и “Качественный состав множества”. В то же время теория К-исчислений обладает несколькими общими глобальными преимуществами по сравнению с перечисленными выше, а также другими известными подходами к формализации содержания ЕЯ-текстов.

Во-первых, модель, построенная в данной книге, представляет в математической форме *гипотезу* об общих внутренних (или ментальных) механизмах формирования сложных структур концептуального уровня (или семантических структур) из первичных единиц концептуального уровня. Ни ЭЛ, ни ТКГ не предпринимают попытки подобного рода. Во-вторых, построенная выше модель формулирует *гипотезу о полной системе квазилингвистических ментальных операций*, т.е. внутренних операций, позволяющих строить концептуальные структуры, выражающие смысл произвольных реальных предложений и дискурсов на ЕЯ, относящихся к любым областям деятельности человека. Другие же известные подходы к формальному описанию содержания ЕЯ-текстов лишь отмечают расширение выразительных возможностей (как правило, языка логики предикатов первого порядка), не выдвигая гипотезы о построении модели полной системы квазилингвистических ментальных операций и не обсуждая эту проблему.

В-третьих, форма описания как языка концептуальных графов в ТКГ, так и логических форм в ЭЛ не является строго математической. Стиль описания концептуальных графов в работе (Sowa 2001) напоминает стиль описания языка программирования, например, языка Паскаль. Это потенциально затрудняет разработку на основе этой теории алгоритмов обработки знаний, совместимых с процедурами вычислительной логики. Набор форм Бэкуса-Наура, используемый

в диссертации Ч.Х. Хуан (Hwang 1992) для описания базового логического синтаксиса, включает выражение *<1-местная-предикатная-константа>:= счастливый | человек | определенный | вероятный | ...* , а также несколько других выражений сходной структуры. Единственный способ избежать употребления многоточий в продукциях заключается в рассмотрении некоторого аналога понятия концептуального базиса, введенного выше.

Дополнительными общими преимуществами аппарата СК-языков по сравнению с ТКГ и ЭЛ являются: (1) более четкое структурирование предметных областей на основе определения множества типов, (2) рассмотрение отношения совместимости на множестве сортов предметной области, позволяющее связывать со многими сущностями не одну, а несколько "координат" по разным "семантическим осям", (3) наличие средств формального описания смысловой структуры таких дискурсов, в которых есть ссылки на смысл предыдущих фраз или более крупных частей текста, (4) возможность моделирования смысловой структуры фраз с прямой и косвенной речью, (5) возможность рассмотрения информационной единицы, соответствующей слову "понятие", что расширяет арсенал средств формального представления энциклопедической информации, (6) возможности использования логических связок "и", "или" для соединения обозначений объектов или понятий или целей, (7) возможность использования в формулах имен функций, аргументами и/или значениями которых могут быть множества объектов или понятий.

Наряду с общими преимуществами по сравнению с ТКГ и ЭЛ, теория СК-языков обладает рядом индивидуальных преимуществ по сравнению с каждым из этих подходов. Основными дополнительными преимуществами по сравнению с ТКГ являются (1) возможности построения составных обозначений целей, команд, (2) значительно большие возможности построения составных обозначений множеств объектов и множеств понятий. К дополнительным преимуществам по сравнению с ЭЛ, в частности, относятся возможности построения составных обозначений понятий и множеств объектов.

Примеры текстов, анализируемых в публикациях по теории представления дискурсов (ТПД), по своей сложности близки к сложности предложения "Если у

человека есть автомобиль, то он является владельцем кредитной карты". Поэтому, по всей видимости, причины широкой популярности ТПД являются не столько научными, сколько психологическими: простота восприятия изображений, состоящих из блоков с простыми формулами, импонирует широкому кругу лингвистов, интересующихся семантикой ЕЯ. Что же касается ценности ТПД для приложений, то можно согласиться с мнением профессора Л. Аренберга (Ahrenberg 1992, с. 7) о том, что "вопреки своему названию, ТПД, в сущности, может быть охарактеризована как формальная семантика для коротких последовательностей предложений, но не как теория дискурсов". В отличие от ТПД, предложенная в данной работе модель применима к построению СП широкого многообразия дискурсов на ЕЯ, целей и действий систем, построению определений понятий. Каждое из перечисленных выше преимуществ построенной модели по сравнению с ТКГ или ЭЛ является преимуществом модели по сравнению с ТПД.

Важным аспектом расширения в ЭЛ выразительных возможностей логики предикатов является введение специального оператора *Ка* для образования *видов действий*. Например, выражение "красить стену" в работах (Hwang и Schubert 1993a; Schubert и Hwang 2000a) рассматривается как обозначение вида действия. Поэтому с помощью оператора *Ка* строится выражение *Ка (красить стена)* , интерпретируемое как терм языка логики предикатов. Затем образуется выражение

(настоящ-время [Джон любит (*Ка (красить стена))]* .

Это выражение, с одной стороны, интерпретируется как СП предложения " Джон любит красить стену". С другой стороны, в указанных работах это выражение рассматривается как аналог некоторой формулы языка логики предикатов первого порядка.

Попытаемся ответить на вопрос, насколько математически корректным является выбранный способ расширения выразительных возможностей языка логики предикатов первого порядка (ЯЛП1) с помощью оператора Ка.

С одной стороны, неопределенные формы глаголов с зависимыми словами позволяют выражать цели интеллектуальных систем, назначения вещей (например, "подъемные краны нужны для того, чтобы поднимать, перемещать и

опускать грузы”), советы, желания, умения и т.д. Кроме того, к этому виду выражений легко привести и императивные предложения (в частности, команды).

С математической точки зрения, оператор Ка в указанных работах рассматривается как функция, одним аргументом которой является предикат “Красить”, а другим – терм (информационная единица, соответствующая слову “стена”). Но одно из принципиальных ограничений логики предикатов первого порядка заключается в том, что не расматриваются функции, аргументом которых могли бы быть предикаты. Кроме того, выражения, образованные неопределенными формами глаголов с зависимыми словами, могут быть сколь угодно сложными и длинными. Чтобы учесть все возможные случаи, нужно рассматривать счетное множество функций, аргументами которых могут быть как предикаты, так и термы.

Представляется, что такой подход полностью противоречил бы как нашей языковой интуиции, так и принципам математической логики. Именно поэтому в работах по ЭЛ рассматриваются только очень простые выражения подобного вида. Как правило, это неопределенная форма глагола с зависимым существительным.

Между тем, в теории СК-языков предлагается такое оригинальное решение этой проблемы, которок полностью соответствует нашей языковой интуиции и не является насилием над принципами какого-либо математического подхода. Это решение заключается в возможности построения сколь угодно сложных выражений, интерпретируемых как семантические представления выражений, образованных инфинитивами вместе с зависимыми словами. Например, цель “Поступить в Московский государственный университет им. М.В. Ломоносова, окончить его с отличием, подготовить и защитить кандидатскую диссертацию по физике” можно представить в виде К-цепочки

*поступление1 * (Учеб-заведение, нек университет * (Название, ‘МГУ им. М.В.Ломоносова’) : x1) ∧ окончание1 * (Учеб-заведение, x1) ∧ подготовка1 * (Объект1, нек диссертация * (Вид, кандидат)(Область1, физика) : x2 ∧ защита1 * (Объект1, x2)) .*

4.15. Обсуждение построенной математической модели

Многочисленные примеры, расмотренные в этой главе, показывают, что выразительные возможности СК-языков соответствуют постановке задачи в параграфе 3.1. Анализ литературы показывает, что разработанная модель для описания структурированных значений (СЗ) ЕЯ-текстов обладает целым рядом отличий от известных формальных и “полуформальных” подходов к описанию смысловой структуры текстов на естественном языке.

Представляется заслуживающим внимания обсуждение вопроса не только о содержании разработанной модели (т.е. о характере предлагаемых моделью операций на СЗ текстов), но и о форме новой модели. Поиск адекватной формы модели был сопряжен с рядом трудностей. Во-первых, на определенном этапе исследования стало ясно, что модель должна представлять собою индуктивное определение, задающее одно вспомогательное и десять основных правил построения формул. Однако получавшееся индуктивное определение было слишком сложным для восприятия, поскольку требовало одновременного понимания всех одиннадцати способов построения формул трех видов с учетом взаимодействия этих способов.

Поэтому в данной работе предлагается шаг за шагом задавать расширяющиеся множества формул $Forms_k$, где k = 0, 1, …, 10, определяемые совместной индукцией некоторыми правилами P[0], P[1], …, P[k] . Форма каждого из правил P[k] где k = 0, 1, …, 10, такова, что правило P[k] может использоваться без изменения в определениях множеств $Forms_k$, $Forms_{k+1}$, ..., $Forms_{10}$. Такой подход позволяет детально проиллюстрировать эффект введения нового правила P[k] , где k = 1, …, 10, с точки зрения расширения множества определенных на предыдущем шаге формул $Forms_{k-1}$.

Во-вторых, в логике предикатов отдельно определяются сначала множество термов, а затем множество формул, причем составные термы интерпретируются как обозначения различных объектов. Но мы знаем, что в русском, английском и многих других языках для построения составных обозначений объектов могут, в частности, использоваться причастные обороты и придаточные определительные предложения. Как следствие, структура таких составных

обозначений объектов различных видов может быть ничуть не проще, чем структура фраз, выражающих высказывания. Поэтому в данной работе предложено в одном определении задавать множество формул, часть из которых может интерпретироваться как составные обозначения объектов (т.е. как термы).

В-третьих, подход к концептуальномк структурированию предметных областей, предложенный в данной работе, существенно отличается от подхода логики предикатов. В логике предикатов исходят из того, что есть некоторые объекты, функции, заданые на объектах, и высказывания об объектах. Множество объектов не структурируется в традиционной логике предикатов и делится на классы (сорта) в многосортных логиках. Функции не могут быть определены на высказываниях (кроме функции, задающей истинностное значение высказывания), и их значениями не могут быть высказывания.

Между тем, на высказываниях могут быть определены различные функции, например, *Автор (Авторы)* и *Дата*. С другой стороны, естественным было бы рассматривать и функции, значениями которых являются высказывания. Например, если сущность является понятием, то значением функции *Определение* может быть формула, поясняющая смысл понятия.

В данной работе предложен новый подход к концептуальному структурированию предметных областей. Образно говоря, этот подход обладает значительно большей *"разрешающей способностью"* по сравнению с подходом логики предикатов. Это увеличение разрешающей способности структурирования действительности в данной работе обеспечивается:

(1) формальным различением (а) обозначений объектов и обозначений понятий, характеризующих эти объекты, (б) обозначений сущностей (предметов, событий, понятий) и обозначений множеств, состоящих из этих сущностей;

(2) наличием средств формального представления упоряоченных наборов и множеств, состоящих из упорядоченных наборов (т.е. *n*-арных отношений);

(3) возможностью рассматривать функции, аргументами и/или значениями которых могут быть семантические представления ЕЯ-текстов, множества объектов, множества понятий.

В свою очередь, для реализации перечисленных выше идей потребовалось задавать совместной индукцией три класса выводимых формул, а затем каждой формуле одного из этих классов (множества *Ls(B)*, где *B* - концептуальный базис) поставить в соответствие цепочку, называемую типом данной формулы и интепретируемую как концептуальную характеристику сущности, обозначаемой рассматриваемой формулой.

Таким образом, предложенный метод пошагового ввода правил построения формул с последующим соединением этих правил с помощью итогового индуктивного определения является оригинальным и может рассматриваться как один из научных результатов данной работы. Этот метод обогащает дискретную математику (в первую очередь – теорию формальных языков) и может использоваться для создания новых исчислений, формулы которых предназначены для описания структуры сложных линейных объектов.

В последние три десятилетия класс языков логики предикатов первого порядка (ЛППП) являлся стандартом, с которым сравнивались предлагавшиеся новые подходы к формальному представлению содержания ЕЯ-текстов. Чаще всего, такие новые подходы рассматривались их авторами как расширения языка ЛППП. Учитывая это, можно выделить следующие преимущества класса стандартных К-языков по сравнению с классом языков ЛППП (при этом будет использоваться нумерация свойств ЕЯ-текстов, введенная в параграфе 3.1):

(Св. 3) возможность строить и формально различать обозначения единиц, соответствующих (а) объектам, ситуациям, процессам в реальном мире и (б) понятиям, квалифицирующим (характеризующим) эти объекты, ситуации, процессы; (Св. 4.1) возможность строить и различать обозначения объектов и множеств объектов; (Св. 5) возможность различать формальным образом понятия, квалифицирующие объекты, и понятия, квалифицирующие множества объектов тех же видов;

(Св. 6) возможность строить составные обозначения понятий, т. е. строить формулы, отражающие поверхностно-семантическаю структуру ЕЯ-выражений, подобных выражению “человек, окончивший МГУ имени М.В. Ломоносова и являющийся биологом или химиком”; (Св. 8) возможность строить обозначения упорядоченных n-местных наборов различных сущностей, где $n > 1$;

(Св. 9). возможность строить (9.1) формальные аналоги составных обозначений множеств, (9.2) обозначения множеств упорядоченных наборов сущностей (9.3) обозначения множеств, состоящих из множеств;

(Св. 11) возможность моделировать смысловую структуру фраз, содержащих, в частности: (11.2) выражения, полученные применением связок “и”, “или” к обозначениям (11.2а) предметов, событий; (11.2б) понятий; (11.3) выражения , где связка “не” стоит непосредственно перед обозначением предмета, события, понятия и т. д.; (11.4) косвенную речь; (11.5) причастные обороты и придаточные определительные предложения ;

(Св. 14) возможность моделировать смысловую структуру дискурсов со ссылками на смысл фраз и более крупных фрагментов рассматриваемых текстов содержит семантическое представление дискурса “У А.Зубова есть три друга. П.Сомов это знает”; (Св. 17) возможность рассматривать нетрадиционные функции (и другие нетрадиционные отношения) с аргументами и/или значениями, являющимися множествами предметов, ситуаций.