Вестник РГГУ, 2009, № 6

Н.Н. Леонтьева

О семантике – в Лаборатории и после

Дискуссия о семантике в теоретической и прикладной лингвистике. Лаборатория машинного перевода. Элементарные смыслы. Проблема «стыковки» лингвистических и «предметных» знаний. Информационно-лингвистическая модель (ИЛМ) и ее отличия от модели «Смысл-Текст» (МСТ).

Ключевые слова: машинный перевод, элементарный смысл, семантический анализ, смысловые отношения, предметные знания, информационно-лингвистическая модель, «мягкое» понимание.

50 лет назад в стенах Лаборатории машинного перевода (ЛМП) 1 МГПИИЯ была заложена теория «элементарных смыслов». Одно из научно-прикладных направлений ЛМП привело к созданию информационно-лингвистической модели (ИЛМ) – с более крупными единицами понимания текста (от «элементарных ситуаций» до «текстовых фактов»). Дальше об этом будет говориться подробнее.

Сейчас, по прошествии многих лет, я бы остановилась на одном аспекте споров и противоречий, связанных со словами *Смысл* и *Семантика*, поскольку до сих пор нет единого толкования стоящих за ними понятий. Лингвисты считают их своими терминами. Действительно, семантика – это сердцевина лингвистики. Но две неразлучные дисциплины, они же родные сестры – теоретическая лингвистика (ТЛ) и прикладная лингвистика (ПЛ) – часто не могут придти к согласию в вопросе о том, что такое семантика, где она начинается и где кончается. Прикладная лингвистика вынуждена принимать во внимание иное, чем в ТЛ, более узкое и более специальное понимание термина и понятия «Смысл». Ведь назначение ПЛ, а точнее, систем автоматической обработки текстов – обслуживать не только и не столько лингвистов, сколько любых специалистов из других предметных областей, нуждающихся в быстрой обработке текстов на естественном языке (ЕЯ).

У лингвистов и «специалистов» (не-лингвистов), работающих с ЕЯ-текстами, это выливается в создание разных моделей. Серьезные лингвисты стремятся к созданию фундаментальных моделей, полностью соответствующих теории языка

и реализуемых на компьютере. А «специалисты», вынужденные часто создавать **свои** модели прикладных систем анализа профессиональных текстов, называют их «моделями на выброс», признавая их экспериментальный и временный характер. Можно ли примирить эти два подхода и найти общее решение проблемы автоматического понимания текстов на едином для тех и других естественном языке? Можно ли придти к единому толкованию слова *Смысл*?

По-настоящему, системно, решить это противоречие может только серьезная теория. Но я склоняюсь к тому, что искать решение надо даже не в ТЛ, а в обобщении опыта создания систем **автоматического понимания текстов** (АПТ): ведь этот опыт и позволяет проверить практикой адекватность тех или иных определений спорных понятий. В компьютерных системах, имеющих дело с текстами на ЕЯ, резче обозначены границы между уровнями, поскольку разным уровням соответствует различный формальный аппарат, и приходится на практике преодолевать или как-то иначе решать проблему «стыков» – между лингвистическими уровнями и не только. Вот на таких стыковочных участках положения ПЛ порой оказываются «теоретичнее» ТЛ, особенно когда надо привлекать к анализу внетекстовые знания, без которых любое определение Смысла текста будет неполным, или договориться о том, что такое адекватность понимания.

Однако мало кто из лингвистов готов согласиться с тем, что такие участники процесса понимания, как пользователь, программист, логик или «специалист» в какой-либо узкой предметной области могут внести уточнения в коренные вопросы собственно лингвистики. Но и «специалистов» не всегда можно убедить, что без серьезной лингвистической основы проблему интеллектуальной обработки знаний решить не удастся. «Специалисты», создавая системы АПТ, вынуждены «сражаться» с лингвистическими проблемами сами, кто как умеет. Лингвистам же трудно апеллировать к «специальным» знаниям, которых слишком много и они разные, а еще труднее «исключить» их из текстов: аппарат не позволяет. Тем не менее, именно лингвистам надлежит определять, что такое **Смысл** и как его фиксировать, а остальным специалистам – уметь искать **Смысл** в текстах, опираясь на научно обоснованные понятия. Пока оказывается, что это не один и тот же объект.

Теория систем АПТ еще не сформировалась, но у нее уже непростые отношения с лингвистической теорией семантики. Поскольку дискуссия по этому вопросу не только не закрыта,

но практически даже не начиналась, я и осмелилась вынести ее на обсуждение – с большой надеждой, что проблема имеет положительное решение. Итак, это проблема «стыка» ТЛ и ПЛ, но лишь в том понимании (допускаю, что оно ограниченное), которое сформировалось у меня как результат личного опыта работы с прикладными системами, начиная с участия в разработке простых моделей машинного перевода на основе «элементарных смыслов» (1959 год, ЛМП) и заканчивая сегодняшним днем – периодом развитых информационных интересов и технологий (2009 год).

1. Об элементарных смыслах

Многим (из разных публикаций и учебников) известно, что именно в ЛМП начались работы по поиску «элементарных смыслов» (ЭС), как определил Вячеслав Всеволодович Иванов задачу Лаборатории в 1958 году. Конечно, В.В. Иванов следил за развитием работ по машинному переводу (МП) в США и Европе, проблема МП вызвала настоящий бум в то время. И не на пустом месте возникла идея семантических примитивов, а в подражание или развитие работ по информационным системам, прежде всего Перри, Берри и Кента, которые называли их «семантическими множителями»[1]. Этот последний термин стал на долгое время символом (паролем, знаменем) ЛМП.

Что касается идеи элементарного смысла, сошлюсь на статью трех авторов, начавших эту работу (Жолковский-Леонтьева-Мартемьянов), опубликованную в сборнике «Машинный перевод» в 1961 году[2] и перепечатанную в скорректированном виде в книге Ю.С. Мартемьянова[3]. Главная мысль статьи состояла в том, что «Задачей перевода является передача на другом языке не грамматики, а смысла переводимого текста». Это было как бы противопоставление недавно вышедшему переводному (с английского) сборнику «Машинный перевод»[4], где больше обсуждались синтаксические и технические проблемы (недостаток памяти для словаря и др.). В то же время необходимо отметить, что уже в этом сборнике были подняты и проблемы неоднозначности (статья В.Ингве), и несколько важных семантических проблем (например, намечено понятие семантической роли, не меняющейся при синтаксических преобразованиях, и даже дано определение семантического согласования – в статье Л. и А. Вундхейлеров).

Напомню основные сведения о метаязыке ЭС и структуру записи на нем, как они сложились в ЛМП и были отражены в упомянутой статье.

Семантическая запись простого предложения имела вид P(A,B,C), что соответствует форме многоместного предиката с его сильными аргументами (реально больше 3-х аргументов в нашем эксперименте не встретилось). Если в предложении была «слабая» группа (Д), она присоединялась внешним, опоясывающим предикатом: E(P(A,B,C),Д), где E – смысловой эквивалент связи. Каждая из лексических единиц, включая сам главный предикат, представлена в записи некоторой комбинацией ЭС – так, как задан (определен) «смысл» данной лексемы в семантическом словаре. Синтаксис ЭС был стандартный: определяемое после определяющего, а на N-е место предиката должны встать его зависимые. Так, слово *понимать* имело смысловой код 22.25.0(31.26,), оно расшифровывается как ЭС «умственность-22»+«восприятие-25» одушевленного субъекта («одушевленность-31» и «элемент»-26). Знак 0 использовался для обозначения двустороннего отношения «исходности»: выражение *A имеет B* запишется как 6.0(A,B), но *B принадлежит A* – как 0.6(B,A). Для выделения основного члена ситуации введен был знак подчеркивания:

Собственность	6.0 (31.26, <u>B</u>)
Владелец	6.0 (<u>31.26</u>, B)
Принадлежать	<u>0.6</u> (B, 31.26).

Авторы концепции ЭС не придавали ей статус общелингвистической теории. Напротив, заявлено, что разложению на ЭС разумно подвергать лишь те лексемы, которые активно участвуют в языковом перифразировании (*дать-получить-иметь; все-каждый*). Конкретные предметные слова, даже и предикатные, исключались из рассмотрения. Хотя уже с первых строк статьи объявлялось, что «Важны, в первую очередь, …смысловые отношения между смысловыми же компонентами высказывания», собственно «смысловые отношения» появляются лишь как значение связи со слабой группой или придаточным предложением. Неявно, парадигматические сходства и различия между понятиями можно было видеть в частично совпадающих наборах ЭС; так, *зона* описана в словаре двумя множителями: 20.26 («пространственный элемент»), а *страна* – тремя: («пространственный одушевленный элемент» – 20.31.26), где 20=«пространство», 26=«элемент»; они различаются на ЭС 31= «одушевленность».

Таких ЭС было всего 3 десятка, с их помощью можно было представить значения около 100 слов, содержащихся в небольшом политическом тексте из журнала Moscow News, издававшегося на пяти языках. Нами были созданы 3 алгоритма

анализа текстов на английском (А. Жолковский), французском (Ю. Мартемьянов) и шведском (Ю. Щеглов) языках (анализы кончались построением одинаковой семантической записи) и алгоритма независимого синтеза, со всеми необходимыми трансформациями, на русский язык (Н. Леонтьева). Работа по программированию всех алгоритмов проводилась в ИТМ и ВТ (Институт точной механики и вычислительной техники АН СССР). Реализация затруднялась внешними обстоятельствами. Однако некоторые выводы можно было сделать уже на этапе подготовки к эксперименту.

Что касается «перевода по смыслу», то вряд ли кто будет отрицать важность этого призыва, но до его реализации в устойчивой и воспроизводимой форме пока не хватило пятидесяти прошедших лет. До сих пор осталось еще много нерешенных проблем – в основном лингвистического свойства, но их решают в основном как отдельные частные задачи. Мне кажется, что небесполезно посмотреть на проблему машинного перевода более поверхностным, но и более прагматичным взглядом. И в ней не обойтись без нескольких «аспектов»:

1. Аспект внешней прагматики – кому и зачем нужен МП (помимо чисто научной области применения как сферы проверки правильности лингвистических решений).

2. Аспект внутренней прагматики – без каких компонентов / крупных «деталей» не может обойтись система МП.

3. Аспект выбора внутреннего инструментария – какой способ представления смысла приведет скорее к цели (при том, что целью всегда остается «перевод по смыслу»).

Проблема «элементарных смыслов» (ЭС) относится к третьему аспекту, т.е. выбору инструмента анализа естественных текстов. Я начала с нее – не только потому, что она больше двух других проблем связана с ЛМП. В эксперименте с моделью ЭС прагматика подразумевалась в самом общем виде: МП был востребован обществом, а сама задача была увлекательной с лингвистической точки зрения. Но ведь характер метаязыка полностью зависит от тех задач, которые он призван обслуживать, т.е. от того, как заданы аспекты 1 и 2: они определяют и всю стратегию анализа, и круг привлекаемых к анализу средств (в первую очередь словарей). Ниже я к этому вернусь (см. п.4).

О том, как уточнялась со временем задача МП и какой вклад внесли в эту проблему работы ЛМП, я могу говорить только от первого лица, то есть достаточно субъективно. Главное состоя-

ло в том, что впервые был предложен вариант семантического метаязыка, на котором можно было представить в идеале любой текст, а также задать некоторые необходимые при переводе правила смыслового перифразирования. Эти достижения не пропали даром, но работа вскоре приняла несколько другой поворот: включившийся в работу Лаборатории И.А. Мельчук придал ей более строгий характер, сосредоточившись на синтаксических деревьях. Так начался второй важный этап в жизни ЛМП.

2. Появление модели и теории «Смысл↔Текст»

В начале 60-х гг. у нас и за рубежом началась глубокая разработка синтаксического уровня, трансформаций и т.д., на первый план вышли системы МП второго поколения. Работа шла с понятием дерева предложения, его синтаксических связей (синтагма, конфигурация, модель управления и др.), а также неоднозначности в установлении внутрифразовых связей. Вышла книга «Автоматический перевод»[5], отражающая исследования по МП почти всех ведущих коллективов мира. В ней основное внимание было направлено именно на явление неоднозначности, которая не решается только средствами синтаксиса. К тому времени И.А. Мельчук[6] и Л.Н. Иорданская[7] уже проделали грандиозную работу по описанию русского синтаксиса в алгоритмическом виде. И.А. Мельчуком и О.С.Кулагиной было создано несколько версий систем машинного перевода[8], с английского и французского языков на русский. Мельчук, будучи фактическим руководителем нескольких научных направлений, в том числе и в ЛМП, первым переключился серьезно на проблемы семантического словаря и «смыслового» синтеза. Лаборатория кипела от горячих споров вокруг темы: «Где взять тот самый смысл, с которого можно строить выходной текст на основе смыслового же словаря?».

Поиски адекватного определения смысла продолжались в ЛМП в жанре описания семантики групп слов. При этом давались не только толкования лексем, входящих в какую-либо тематическую парадигму (время, целесообразная деятельность, воля, сила, имущественные отношения и др.), но и формулировалась вся аксиоматика, поддерживающая определения слов в возможном семантическом словаре. Большая часть этих работ нашла отображение в сборнике МПиПЛ, выпуск 8, вышедшем в 1964 г.[9] Сборник был переведен в дальнейшем на английский

язык и получил широкий международный резонанс. Хотя этот цикл работ ЛМП достоин большего внимания, я не останавливаюсь на нем, так как это все же в основном словарные работы, и их лучше освещать при более серьезном и специальном описании именно семантических словарей нового типа. (Правда, Ю.С. Мартемьянов пытался формализовать запись ситуаций для небольших связных текстов, но это направление не стало общим, а в дальнейшем оно было описано в его книге, см. ссылку 3).

Возник тандем Мельчук-Жолковский, этот творческий союз подарил миру идею и разработку лексических функций (ЛФ)[10]; их основой могла быть собираемая в то время в ЛМП картотека «пустых» слов, которые трудно было описать в категориях ЭС. Понятие лексической функции как вида типовой лексической связи стало ценным аппаратом описания части лексем и словосочетаний с «трудной» семантикой. Конечно, введение ЛФ не могло компенсировать полностью отсутствие «грамматики смыслов» в теории ЭС. Развитие экспериментальной модели ЭС на этом прекратилось, а разные варианты списка ЭС используются до сих пор как вспомогательное средство различения значений, дополнительно к синтаксическим свойствам слов и конструкций. Имея синтаксические структуры как твердое начало (и даже синтактико-семантические, поскольку в синтаксическом дереве появились сложные семантические узлы вида «ЛФ + ее аргумент»), можно было выходить на новый уровень, включающий семантические толкования лексем. Но на этом шаге произошла длительная остановка. Семантические толкования лексем, как они отражены в МПиПЛ №8, были слишком сложными и одновременно «размытыми», далекими от возможности их формализации и тем более от машинной реализации.

Основным направлением ЛМП стала работа над моделью, получившей имя «Смысл-Текст» (МСТ)[11]. В разработку включился Ю.Д. Апресян, в дальнейшем МСТ приобрела статус «Теории Смысл-Текст», или ТСТ, с тремя авторами: Мельчук-Апресян-Жолковский. Описание значений лексем приобрело системный вид. Вторым после синтаксиса необходимым компонентом этой теории и модели был объявлен толково-комбинаторный словарь (ТКС)[12]. Описание лексем по формату ТКС (для русского, английского и французского языков) стало на долгие годы столбовой дорогой не только работ в ЛМП, но и всей отечественной структурной лингвистики.

3. Другие направления в ЛМП. Интерес к «целому тексту»

Параллельно с ТСТ в лаборатории развивались и другие, боковые течения. Так, продолжали работать с художественным текстом Жолковский и Щеглов[13]; Мартемьянов и Г.В. Дорофеев исследовали логические связи и естественные выводы в коротких текстах[14]; С.И. Гиндин работал с лингвистикой «целого текста»[15], Э.И. Королев стал заниматься информационно-поисковыми системами (ИПС), были и другие увлечения. Наиболее серьезный фронт работ развернулся вокруг возникшей на новой основе задачи англо-русского автоматического перевода (АРАП), которую поддерживал недолго хоздоговор с ВЦ Армянской ССР. В рамках этой системы создавался англо-русский многоаспектный автоматический словарь (АРМАС), руководство которым осуществляла З.М. Шаляпина[16].

Мой интерес к информационным системам подогревался двумя обстоятельствами: в них был и целый текст и даже массивы текстов с их прагматикой (запросы и вопросы пользователя), но там был и лингвистический анализ. Было нелегко сделать выбор между двумя крайностями: глубокой лингвистикой, имеющей дело с уровнем предложения (линия Мельчук, Апресян), и лингвистикой как глубоко прикладной, практической дисциплиной. Это последнее диктовалось хоздоговорами (начиная с Главного ВЦ Госплана СССР как заказчика работ и потом других). Мне по разным причинам выпало второе, о чем и сейчас не жалею, но много сил ушло в дальнейшем на попытки соединить это второе с первым, в чем я не встретила поддержки ни с той, ни с другой стороны. К тому же оказалось, что научные расхождения (разные взгляды на одну проблему и способы ее решения) тесно связаны с человеческими отношениями. Чтобы «развести» их, я сделала не одну попытку показать неконфликтность и даже дополнительность двух установок: одной – на детальный лингвистический анализ предложений, другой – на создание информационного образа целого текста.

В итоге я пришла к выводу, что необходимо строить более прочный мост, соединяющий лингвистическую семантику с информационными системами (далее ЛС и ИС), что нужно теоретически обосновать эту связь (в частности, уточнить, как соотносятся понятия «Информация» и «Смысл»). Вряд ли можно называть результаты первой установки (ЛС) объективными, а результаты второй субъективными, как иногда их интерпретируют. В обеих есть изрядный процент субъективизма, но в ин-

формационной системе субъективизм принципиален, так как ИС обязательно включает адресата, а каждый адресат ищет свою (субъективную) информацию. Формально это противоречие выражается в том, что лингвистический анализ стремится к построению единственной правильной структуры, к единственной «истине», а ИС должны строить столько разных информационных структур, сколько будет вопросов к тексту, ибо моделируют разные понимания одного источника. Но ведь разные понимания одного объекта – одно из главных свойств человеческого восприятия. И создать модель, строящую разные результаты понимания вербального материала естественных текстов, – очень увлекательная задача.

В последнее время необходимость соединения методов информационного и лингвистического анализа в один работающий комплекс подтверждается практикой многих зарубежных систем. Существуют методы и работающие программы, извлекающие из текстов заказанную информацию, причем относящуюся не только к объектам с их параметрами, но часто и к действиям над ними. Это системы типа Information Extraction (IE-системы), обычно ограниченного масштаба и настроенные на одну предметную область. В таких задачах заранее определены и формат создаваемых баз данных (БД), и лексика, которая может заполнять поля БД. Поиск ведется в заранее отобранном (информационной системой) массиве, где заведомо содержатся искомые сущности. Для их выявления и проводится лингвистический анализ. Но в IE-системах, в отличие от отечественных подходов, слабо развит как раз лингвистический компонент. Хотя они и работают с «целым текстом», но ищут ограниченный круг объектов с их параметрами (в основном имеющими количественное измерение), для которых жестко заданы лингвистические способы оформления. В развитых странах полагаются скорее на информационные технологии, чем на поиск более содержательных взаимодействий текста с компьютером.

Повышение **содержательности поиска** нужной пользователю информации по тексту или корпусу текстов остается в основном за лингвистикой. Она должна идти «навстречу» задаваемым в предметных областях описаниям и уметь строить более естественные и более крупные единицы, чем узлы синтаксического дерева, выраженные преимущественно отдельными лексемами. Ведь стыковка текстовых структур с любыми другими вербальными структурами предметных областей – все же лингвистическая задача: она должна обеспечить переводи-

мость их содержания с языка-1 на язык-2, даже если эти языки принадлежат одному естественному (ЕЯ).

4. Нужны ли «другие» лингвистические модели?

Начав в предыдущем разделе разговор об информации и семантическом метаязыке, я практически перехожу к обоснованию собственной научной линии, тем более что другие подходы достаточно полно отражены и продолжают отражаться в публикациях. Находясь на стыке текста и предметных областей (ПО), проблема их взаимодействия если и затрагивала лингвистов, то скорее в модальности «Отстаньте». Это всегда меня удивляло, так как семантика текста неизбежно выходит «в жизнь», только не прямо, а «вербально». Как это реализовать в системе АПТ? Ждать полного формального описания семантики всех ПО? Или заказать концептуальную схему, пригодную для всех наук и дисциплин? Это нереально. Значит, остается единственный путь – **сравнение двух типов «текстов»**: с одной стороны – сам анализируемый текст, с другой стороны – естественные источники, уже имеющиеся в той ПО, которая вовлечена в процесс АПТ. Это своеобразный машинный перевод с общего языка на специальный. Он обостряет интерес к лингвистическим моделям анализа текста: справятся ли с таким МП? Казалось, что в этой задаче невозможно обойтись без какого-то метаязыка, объединяющего общезначимые и специальные тексты.

Выполняя в ЛМП работы по синтезу (с трех названных выше языков), я сделала для себя такие внутренние выводы: а) лексемы не нужно рассыпать на атомы, так как не определен синтаксис самих ЭС, а их набор и даже имена субъективны и вызывают много споров; б) напротив, семантику связей аргументов со своими предикатами нужно задавать в явном виде **элементарных смысловых отношений** (ЭСО), они и будут служить атомарными формулами в семантическом представлении фразы и, далее, всего текста. Таких единиц в языке намного меньше, чем ЭС, к тому же, их легче сформулировать и найти переводы на другие ЕЯ. Само множество ЭСО может быть набрано для любого языка, если все связи, вычленяемые в этом языке, заменить словами и затем провести работу по обобщению сходных по семантике слов и минимизации их количества. Иначе говоря, надо лексикализовать все виды бинарных грамматических и синтаксических связей. Мы придем к очень похожим спискам, которые можно свести к единому списку, совместимому со всеми

исходными. В этом убеждает и практика создания международного языка UNL первой версии[17], который на 90 % совпал со списком ЭСО, созданным ранее как обобщение семантики русских предлогов[18]. Элементы такого единого списка и будем считать **метасловами**, единицами семантического метаязыка (пока не будет предложен другой метод).

Следующая моя статья уже описывала этот новый подход (см. статью 1967 г.)[19]. Лишь через несколько лет после этого я случайно узнала о статье Ч. Филлмора (1968 г.) о семантических падежах[20], что прибавило уверенности в полезности найденных семантических реалий.

Но я все же не стала пользоваться понятием «падеж», так как ЭСО присутствуют не только как характеристики связей лексемы-предиката со своими зависимыми, но они могут соединять также содержательные единицы по всему тексту. Это и значение любой – сильной или слабой – синтаксической связи, и связи между предложениями, и связи между содержательными частями слова, и между всем текстом и его заголовком и т.д.; они же могут связывать и сами ЭС. Таким образом, ЭСО оказывались новой универсальной единицей описания значений в тексте. Но они оставались при этом вполне элементарной единицей – кванта информации, кванта сообщения: P(A,B) – о единице A говорится, что она находится в отношении P к другой единице B, что можно уже считать **элементарной ситуацией** лингвистического плана. Поддерживало и то, что в ИПС, получивших тогда широкое развитие, и в их главном инструменте (Тезаурусе) тоже использовалось подмножество подобных ЭСО (род-вид и др.). Оставалось только определить более или менее устойчивый список ЭСО, пригодный для описания синтагматики и парадигматики любых текстов, общезначимых и специальных. (После многих экспериментальных проверок я могу утверждать, что список из примерно пятидесяти основных ЭСО оказывается достаточным для описания семантических связей в любом тексте, а активно используемых отношений вдвое меньше).

Я была уверена, что ЭСО не только не противоречат синтаксическим связям в МСТ, но продолжают их на семантическом уровне. Вслед за моими учителями я тоже стремилась к построению единого, однозначного и непротиворечивого синтаксического представления (СинП) и далее семантического представления (СемП). Считала, что семантика даже с таким небольшим багажом, как Список смысловых отношений и предсказываемые ими семантические классы возможных чле-

нов заданных ЭСО, может справиться с задачей снятия синтаксической омонимии, которая была главным препятствием на пути к получению однозначной структуры. Дальнейшая (с 1976 г.) практическая работа над созданием системы МП[21] изменила эту установку.

Во-первых, я убедилась, что конечно, семантическая интерпретация снимает часть синтаксической омонимии, но добавляет и свою, более серьезную неоднозначность, в том числе для предложений, имеющих однозначную синтаксическую структуру (ср. *Чужой опыт непереносим*, *Сопротивление проводника* и много других примеров). Вторым доводом, хоть и несколько позже, стало изменение политического климата в сторону плюрализма, допускающего разные взгляды на один и тот же предмет, текст и, в частности, на семантическую интерпретацию текста. Вдруг стало ясно, что не нужно так упорно стремиться к одному единственному решению в семантическом представлении: ведь оно-то и должно обеспечивать разные «понимания» одного и того же текста разными реципиентами. Еще одним мотивом, побудившим к раздумью относительно природы СемП, стало знакомство с работами, вводящими фактор прагматики в ткань лингвистических описаний (см. работы Т.В. Булыгиной и др.). Как минимум, в семантической структуре должны быть учтены автор текста и адресат, воспринимающий текст[22].

Из этого следовало, что и в модель понимания текста нужно вводить компоненты, отвечающие за прагматику в системе: Автор, «Воспринимающее устройство (ВУ)», а также компонент, отвечающий за включение внетекстовых знаний хотя бы минимального объема: без них невозможно создание полноценного текстового семантического анализа. Это было легко отобразить на схеме модели. В схему был включен также компонент «Разработчик модели»: к нему должны возвращаться результаты оценки качества перевода и другие рекламации, которые потребуют коррекции словарей и процессоров[23].

Я не видела естественного способа включения компонента знаний, как и активного участника процесса понимания текста, в системы, основанные на синтаксическом подходе. Между тем информационные системы работали с этими компонентами, представляющими «внутреннюю прагматику системы», и это относилось, конечно, к семантике.

Вернемся к «внешней» прагматической установке (кому нужен МП), которая тоже влияет на стратегию анализа текста. Ясно, что ожидать перевода целых связных текстов полностью

автоматически и с хорошим качеством пока нереально. А ведь информация из текстов на неродном языке остро нужна специалистам, хотя бы (и даже предпочтительно) в кратком виде. Недаром возник целый класс систем типа «перевод-реферат». Так абстрактная модель МП расширилась внешней установкой на перевод (и понимание) не только детального содержания исходного текста, но и возможностью передачи части содержания, соответствующей интересам пользователя. **Компрессия содержания текста** стала еще одним аспектом АПТ и нитью, связывающей лингвистику с информатикой. Ведь содержание естественного текста можно сжимать, опираясь на многие его свойства, включая «недостатки» (неполнота и др.). Оказывалось, что такая задача даже более реалистична, чем полный перевод. В итоге схема МП расширилась до информационно-переводческой модели, отражающей разные типы понимания текста[24].

5. Начало проблемы «предметной области» в ЛМП

Еще одно направление работы Лаборатории меньше известно лингвистам других коллективов. Когда в СССР были свернуты, вслед за Америкой, работы по МП и вошли в моду новые термины (АСУ, ИИ, БД и др.), в Лаборатории появился жанр хоздоговоров. Чтобы продолжать работы по системе англо-русского МП и особенно словаря типа ТКС, нужно было зарабатывать деньги. Сотрудников стало гораздо больше (на временные работы приглашались даже программисты), часть переключилась на разработку информационной системы для ГВЦ Госплана СССР; работы продолжались семь лет, я отвечала за этот хоздоговор. В отдельных задачах можно было обойтись достаточно простыми методами обработки текстов (морфологический анализ или индексирование текстов на основе языка словосочетаний в Общем классификаторе наименований разных видов продукции)[25]. Новым в этом направлении работ было то, что потребовалось обратить внимание на явление «предметной области» (ПО), которое до тех пор мало занимало лингвистов.

В Госплан поступали заявки со всего Союза и по разным тематикам, здесь и понадобились мои общезначимые ЭСО[26]. Тексты мы расписывали в терминах ЭСО вручную, как и вопросы к текстам, но интересно было строить алгоритмы их сопоставления и доказывать соотносимость (степень близости) примерно так, как решаются алгебраические уравнения.

Однако на ЭВМ нас не допускали, и тогда я обратилась к заказчикам с вопросом «Сколько же можно плавать в сухом бассейне? – Дайте, наконец, воду!». Нам же не давали не только воду, но часто и зарплату сотрудникам, так как каждый год перезаключался договор, сотрудников всех увольняли и снова зачисляли, а зарплату задерживали, иногда до лета. Тем не менее отчеты нужно было предъявлять каждый квартал. Терпение лопнуло, когда на мое предложение взять двух человек из ЛМП в штат ГВЦ (чтобы вовремя оформляли все дела от заказчика) я получила циничный ответ: «Ну, нет, если мы возьмем Вас, Вы будете авторами системы, а так мы сами».

В поисках нормальных заказчиков мы с В.Ю. Розенцвейгом обошли многие институты (в том числе закрытые), и везде надо было вникать в тематику и потребности и составлять перспективный план работ на несколько лет вперед – все безуспешно. Тогда я и согласилась перейти в отраслевой Институт информации (ЦНИИТЭСтроймаш) заведующим лабораторией ИПЯ и стала, в рамках хоздоговора с ЛМП, заказчиком работ по семантическому анализу текстов. Вместе с Е.В. Урысон[27] мы создавали алгоритмы анализа заголовков к текстам ПО «Строительное, дорожное и коммунальное машиностроение». Они программировались в Гатчине, как и система автоматического индексирования текстов по этой ПО, созданная мною с сотрудниками ЦНИИТЭСтроймаш. Был также создан оригинальный двухступенчатый Тезаурус как инструмент анализа текстов отрасли. В Тезаурусе использовалось разложение терминов на смысловые примитивы, только равные словам ЕЯ, плюс помета семантического класса типа «Устройство-Ус, Действие-Д, Предмет-П» и другие достаточно очевидные (так, слово *Грейдер* = *Машина*-Ус + *Выравнивание*-Д + *Земля*-П»), а алгоритмы индексирования в таких категориях использовали некоторые правила вывода на Грамматике ЭСО[28].

Мне было интересно проверить еще раз возможности аппарата ЭСО на материале и других предметных областей. В ЦНИИ по приборостроению вместе с работавшими там прикладными лингвистами было выполнено описание в форме БД около ста понятий, обозначающих приборы (*амперметр*, *вольтметр*, *зонд* и т.д.), с точки зрения их поведения в тексте (сочетаемость с другими важными для ПО характеристиками). Использовавшийся аппарат я назвала «энциклопедическими функциями (ЭФ)», по аналогии с лексическими функциями, из списка которых тоже были взяты некоторые «энциклопедические» ЛФ. С использованием анкет на таком специфическом,

ограниченном «Приборостроением» языке ЭФ был проведен ручной эксперимент по «автоматическому индексированию» текстов рефератов и проведено сравнение с результатами традиционного индексирования[29].

Третьей областью проверки аппарата ЭСО «в полевых условиях» была тематика сельскохозяйственного машиностроения. Сотрудники редакционного отдела этого института с длинным названием просили помощи у лингвистов, сетуя на неграмотность специалистов, пишущих инструкции по ремонту с/хоз машин («Они пишут *трещины в щелях*», – жаловались редактора). Я предложила им простенький язык «метаотношений» (метаЭСО), позволяющий проверять полноту и «формальную» правильность любого утверждения в тексте инструкций, чем они легко и пользовались. Добавлю, что аналогичный метаязык ЭСО верхнего (по отношению к содержанию) уровня я предлагаю и студентам ТиПЛ на спецкурсах в РГГУ, прежде чем они перейдут к созданию подробного и серьезного СемП высказывания. Этот ЭСО коротко аттестован в моем учебном пособии[30]. А метод анкетного описания понятий, подобный описанию приборов, был развит и применен С.Е. Никитиной в Тезаурусе лингвистических терминов[31].

В дальнейшем метаязык ЭСО был применен также к описанию семантической и композиционной структуры официальных политических документов в экспериментальной системе ПОЛИТЕКСТ[32]. Этот опыт использования созданных лингвистами средств семантического метаязыка для описания семантики разных ПО еще не закончен. Мне он интересен сейчас только возможностью исследования проблемы взаимной адаптации разных систем. А вдруг лингвистика окажется полезной еще и в задаче описания и «мягкого» преодоления стыков в гуманитарной области (скажем, в моделях человеческих и социальных конфликтов)?

Проблема ПО развивалась и дальше – в моих работах с ЛМП по линии семантики.

6. Начало ИЛМ

Через четыре года моей практики руководства «извне» частью семантических работ ЛМП открылся Отдел машинного перевода во Всесоюзном центре переводов (ВЦП), я перешла туда на условии, что ЛМП будет постоянным соисполнителем по договорам. Так я стала заказчиком работ ЛМП на долгий период (вплоть до 90-х годов), а часто оказывалась и их основным

исполнителем. В ВЦП удалось привлечь и студентов ОСиПЛ МГУ для составления словарных статей семантического компонента системы МП. Конечно, работающих на хоздоговоре лингвистов, как и студентов, проблема ПО касалась мало, они были заняты в основном наполнением словарей.

При создании системы ФРАП (французско-русский МП, или АП) первоначальной проблемной областью была математика, так как мы взяли за основу систему перевода французских математических текстов О.С. Кулагиной, вместе с ее словарем французских оборотов. Но в ВЦП, где создавалась в числе других и наша система ФРАП, в его производственной части, шел реальный поток заказов на переводы. Надо было учитывать «требования самой жизни», то есть искать заказчиков и настраиваться на какую-то актуальную предметную область.

Исследование массивов ВЦП (т.е. состава «чемоданов переводов» с французского языка на русский) дало пеструю картину. Было несколько заказов по предметной области «Металлургия», а также по тематике «Развитие пчеловодства в древнем Израиле». Выбрали Металлургию, стали настраивать словари. Все обещания дирекции помочь, обеспечить терминологией и тезаурусами остались обещаниями, пришлось лингвистам разбираться самим, когда выбирать эквивалент *губчатое железо*, а когда *железистая губка*, и т.п.

Когда в системе был реализован этап анализа оборотов и терминов и был сдан соответствующий отчет, пришла разгромная рецензия из Госкомитета по науке и технике: *Почему и зачем металлургия, когда нужна электроника?* Пришлось срочно наполнять словарь новой, другой лексикой, тем более что был найден конкретный заказчик переводов по микроэлектронике (из НИЦЭВТ). Он «материализовался» как компонент ВУ (воспринимающее устройство), совмещая несколько функций: не только заказчик и адресат переводов, но также специалист в какой-то области знаний, а к тому же еще и оценщик качества МП (переводы мы подвергали небольшому постредактированию). Наш непосредственный заказчик передал нам такой диалог с директором НИЦЭВТ, которому он показал наши переводы:

Директор (читает): «Ужасный перевод!»

– Так это же машина переводила!

– Ах, машина?.. (читает еще): «Прекрасный перевод!»

Так была поставлена перед нами первая серьезная научная (она же и жизненная) проблема гибкой адаптации «реальной»

(хоть и строящейся еще) системы МП к разным предметным областям при жестком требовании строить сразу «промышленную» систему. Правда, когда мы сдавали Госкомиссии в 1986 году вторую версию системы ФРАП, я как руководитель настояла на том, чтобы систему принимали в **опытно**-промышленную эксплуатацию, а не сразу в промышленную, как сдавались все остальные 9 систем ВЦП, за что и заработала далеко не первое лишение премий и прочих благ.

Систему О.С. Кулагиной (ФР-2) нам пришлось оставить: она была прямолинейна, а семантика заменялась списками слов, требующих особой обработки. Обдумывание задачи «мягкой стыковки» лингвистических структур со структурами разных ПО, а также ряда связанных с ней задач, заняло не один год. Пришлось не просто вводить уровень Семантического Представления (СемП), но пересматривать границы и функции уровня семантики. СемП текста не могло остаться простой одномерной структурой. Оно должно было отобразить, кроме разнородных аспектов содержания, и возможные варианты сжатия и выхода на разные режимы перевода. Результатом была более универсальная модель понимания текста, названная впоследствии информационно-лингвистической (ИЛМ)[33]. Частично ее функции были реализованы в подсистеме ПОЛИТекст системы РОССИЯ, создаваемой в Институте США и Канады РАН и перенесенной потом в МГУ.

В новой задаче отклонение от столбовой дороги (МСТ или ТСТ) состояло в том, что в ее конструкцию включены свойства информационных систем, прежде всего компонент настройки на предметную область и пользователя, что стало возможным благодаря адаптивным качествам используемого метаязыка ЭСО. Это ослабило требования к синтаксической структуре предложений, которая могла оставаться недостроенной или неоднозначной, так как многие функции передавались семантическому компоненту.

Труднее всего было реализовать банальное утверждение о необходимости вводить Знания в семантический анализ и семантические структуры текста. Ведь абсолютно нереально собрать все или даже замкнутую часть знаний о мире и ввести их в ЭВМ. Чтобы настроить лингвистические процессоры и словари на заданную тематику, требовалось обычно «хирургическое вмешательство» лингвистов и специалистов. Очевидно, что это осуществимо только в небольших экспериментах, но для массового автоматического анализа специальных текстов «хирургическое» введение предметных знаний в состав лингви-

стических компонентов практически невозможно (такой вывод я повторю и сейчас).

А естественный вывод напрашивался сам собой. Ведь все вербальные источники знаний, которые нужно или можно подключить к анализу, – это своего рода тексты, только написанные на «другом» языке (даже если они принадлежат одному ЕЯ). Они должны быть переведены на принятый метаязык, чтобы быть сравнимыми с анализируемым текстом, так как сравнению подлежат только однородные сущности. С точки зрения содержания фрагмент знания представляет собой как бы развернутый вопрос, в нем уже присутствуют все те единицы, в терминах и «в пользу» которых должен проводиться анализ текста. Если речь идет о поиске, они задают контекст, для которого нужно найти максимально близкий фрагмент текста, который и будет служить основанием для ответа. С точки зрения адаптации средств лингвистического анализа к ПО нужно учесть тот факт, что правила перифразирования на принятом метаязыке ЭСО намного проще, чем средства синтаксических или синтактико-семантических преобразований. Общее решение проблемы стыковки лингвистических и специальных знаний может моделироваться на обозримом объекте – семантическом пространстве текста – в режиме исследования его свойств.

7. Семантический анализ в ИЛМ

Сам семантический компонент ИЛМ оказался многоуровневым, он включает построение как минимум четырех структур содержательного уровня. Очень коротко охарактеризую уровни анализа и создаваемые структуры.

Первый уровень – это прямая семантическая интерпретация синтаксических представлений (СинП) всех подряд предложений и построение первичного СемП (СемП1), а еще правильнее – **семантического пространства** (СемПрост) текста (подробнее о нем см. ниже). Второй уровень – операция коррекции первичного СемП: построение более правильных – в соответствии с семантической грамматикой – семантических узлов, постановка некоторых «слабых» актантов на место отсутствующих сильных, замена пустых и местоименных слов их антецедентами и т.п. Здесь же происходит сжатие СемП и укрупнение единиц с целью собрать полную семантическую Ситуацию. На третьем уровне СемП подвергается очередному выравниванию узлов – по результатам сравнения их с внешними единицами, заданными хотя бы в виде Тезауруса, и выявле-

нию основных тематических линий текста. Наконец, четвертый уровень анализа строит информационные структуры – по правилам «вычисления» наиболее информативных единиц. Они могут быть получены либо сравнением исходных словарных «весов» лексем (лингвистический путь), либо сравнением СемП текста с СемП запроса, когда «вес» единиц задан извне (информационный, прагматический путь). Генерируются другие структуры типа «Ситуация», которые и представляют Текст во внешней информационной среде. Можно говорить и о разных степенях редукции структуры при построении ответа на вопросы к данному тексту или при включении основного содержания этого текста в Базы данных и/или знаний.

В прикладной системе мы имеем дело с текстовой, то есть вербальной, средой, а не с действительностью (или жизнью). Чтобы материализовать (вербализовать) соотнесение текстовых единиц с объектами действительности, необходимы лингвистические аналоги для этих объектов. Если нас интересуют факты, события, ситуации, то в качестве их текстовых аналогов уместно ввести соответствующие крупные единицы семантического анализа целого текста: **Ситуацию (Сит), Событие (Соб), Текстовый факт (ТФ)**. Они имеют лингвистический статус, но являются заготовками к построению сложных концептуальных единиц **базы знаний.** Видимо, правильнее говорить о получении лингвистическими единицами **статуса концептуальных единиц**. (Вопрос о том, когда лингвистические единицы становятся концептами, остается пока открытым).

Ситуативное представление (СитП) предложений текста – одна из важных промежуточных семантических структур; СитП задает иерархию лексем, основанную на семантических словарных описаниях, но иную, чем в синтаксическом представлении (СинП).

К входным данным при построении СитП относится результат сегментации текста на простые предложения. (Программа анализа сложных предложений на синтаксическом уровне – их сегментации и затем объединения разорванных частей простых фраз – выполнена, например, Т.Ю. Кобзаревой[34], при этом устанавливаются некоторые связи кореференции; считаю, что использование семантики даже в пределах первичного анализа могло бы уточнить их). Как инструменты построения СитП привлекаются еще два вида ресурсов: а) схема и состав типовой единицы Ситуация; б) значения нескольких полей семантического словаря, определяющие роль, место и вес данной лексемы в структуре Ситуации.

Для простых двусоставных предложений (которых в тексте большинство, учитывая и результаты сегментации сложных) правила построения СитП таковы. Нетерминальный символ СИТ подчиняет лексему, выбранную как лексическое ядро (ЛЯ) ситуации. В стандартном случае им становится главный предикат. Если у него максимальный информационный вес, такое СитП является устойчивым, его ЛЯ переходит в структуру со своими смысловыми валентностями, заполненными иногда на предшествующем уровне. Единица СИТ тоже имеет словарную статью со своим набором валентностей, это перечень отношений, которые могут характеризовать любую ситуацию: ВРЕМЯ(?,СИТ), ЛОК(?,СИТ), УСЛовие(?,СИТ), ПРИЧина(?,СИТ), УТОЧНение(?,СИТ) и другие значения в основном сирконстантных и «слабых» связей. Логическая сумма наборов семантических отношений (СемО), являющихся валентностями единиц СИТ и ЛЯ, образует гипотезы, позволяющие присоединять к этим символам оставшиеся «оторванными» группы слов.

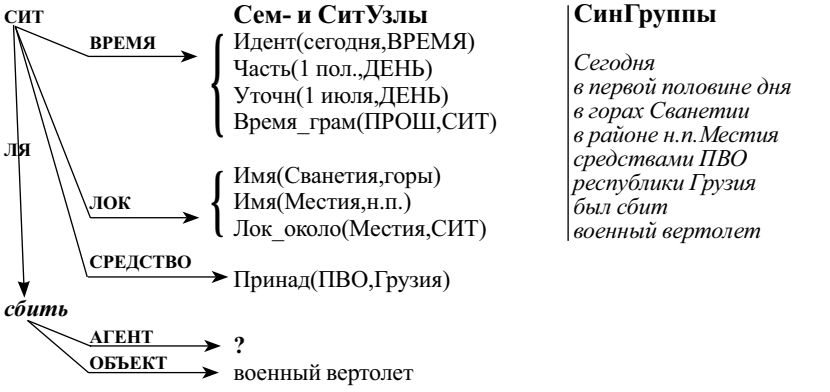
Наибольший вес в словаре имеют лексемы, обозначающие активные действия (*строить*, *разрушать*, *воевать*, *голосовать* и т.п.). У многих предикатов вес зависит от семантики актантов (чаще это смысловой объект или содержание), что фиксируется в нашем словаре. Важны связи модальности и оценки в СитП. Модальность «РЕАЛ» (реальная Ситуация) повышает вес, а модальности, отрицающие само действие, снижают вес до нуля. Некоторые правила построения СИТ и выбора ее лексического ядра приводятся в статье Ермакова М.В.[35]

Если словарный вес главного предиката оказался ниже веса зависимого от него предиката, СитП неустойчиво и требует перестройки иерархий, установленных в СинП. На роль ЛЯ выйдет зависимое; главное слово СинП станет в структуре СИТ зависимым, а его собственная семантическая характеристика будет именем связывающего их СемО. Так, *подготовка к сбору урожая* перейдет в СитП как *сбор урожая* **в стадии** *подготовки*, поскольку ВЕС слова *подготовка* ниже ВЕСа узла *сбор урожая*. Этот последний станет лексическим ядром СитП, а синтаксически главное слово – семантически зависимым; в данном случае именем связи будет собственная характеристика слова *подготовка*:

СТАДия (*подготовка*, *сбор_урожая*).

Ниже приведена упрощенная структура СИТ, первой крупной единицы СемПрост текста, и окончательные информационные единицы, которые могут быть собраны при анализе короткого сообщения из газеты:

1 июля. Сегодня в первой половине дня в горах Сванетии в районе населенного пункта Местия средствами ПВО республики Грузия был сбит военный вертолет.



Информационные Узлы
СОБЫТИЕ = Сбили военный вертолет
АГЕНТ = ПВО Грузии
ВРЕМЯ = 1994 г. – 1 июля
МЕСТО = Грузия – Сванетия – Местия

Статус и формальные свойства следующей по рангу единицы («Событие», «**Соб**») приобретают не все, а лишь семантически сильные единицы Сит. Лингвистический анализ в системе АПТ должен уметь настраиваться на объективное выявление «сильных» Ситуаций: ведь только они станут теми единицами, которые уместно внести в БД «События». Так, узел *подготовка к войне* семантически слабее, чем узел *война* (в модальности РЕАЛ), *помощь при тушении пожара* слабее *пожара*, *замысел убийства* слабее *убийства*, *обсуждение доклада* слабее *доклада* и т.п. По умолчанию Событием (Соб) может быть объявлена Ситуация, если ее лексическое ядро имеет максимальный словарный ВЕС (обозначает активное действие), контекст не привел к снижению веса, модальность СИТ вычислена как РЕАЛьное действие, а ее аргументы реализованы в ближайшем контексте или получили поддержку в тексте.

Сильной единица типа Сит будет и в том случае, если лексическое ядро Сит войдет в списки слов-событий, заданные самим пользователем (субъективное Соб). Например, пользователь захочет собрать сведения из текстов именно обо всех видах помощи какому-то государству независимо от события, которое явилось причиной – *наводнение, землетрясение, пожар* ... Тогда даже слабая единица ЛЯ(Сит) станет субъективно

сильной. Гибкость метаединицы Сит позволяет структуру Сит перестроить «в пользу», или «навстречу», запросу пользователя – тогда тоже получим другой состав единицы Событие, отражающий вид самой помощи.

Для синтаксически неполных предложений будет построено заведомо неполное СитП. Предикат односоставного предложения (например, *Война.*) займет место ЛЯ, а валентности узлов СИТ и ЛЯ останутся пока незаполненными. Многие лексемы и выражения с очевидной семантикой займут соответствующее их основной характеристике место в структуре СИТ, например, односоставная фраза *Осень 1993.* встанет на первое место СемО ВРЕМЯ: ВРЕМЯ(*Осень_1993*, ?СИТ), т.е. на данном отрезке текста остается неясным, о какой ситуации пойдет речь в тексте. Даже в случае односоставных предложений с «предметными» лексемами (*Река. Нефть. Рыба.*) им отведена роль пока неясного Актанта неизвестной СИТ. Выяснение того, к каким ситуациям присоединить такие изолированные узлы, будет отнесено на следующие этапы. Сама фиксация семантически необходимых, но отсутствующих на данном отрезке участников Ситуации создает формальный стимул для перехода к анализу целого связного текста.

Рассмотренные выше единицы СИТ и СОБ имеют локальный характер: они могут представлять содержание отдельного высказывания, абзаца, но не быть важными в контексте целого текста. «Текстовый факт» (ТФ) – это структурная единица, представляющая содержание всего текста. В простом случае ТФ можно собрать как последовательность всех СитП, которые удалось построить при анализе: ТФ(Текста) = СитП1 + СитП2 + …+ СитПк. Такая структура ТФ близка к результату лингвистического анализа. ТФ можно собрать и как последовательность всех Событий, построенных на основе множества ситуативных представлений: ТФ(Текста) = СОБ1 + СОБ2 + … + СОБк, и это будет более сжатая, но и более далекая от лингвистической структура. Оба эти вида ТФ собирают «действия» или близкие к ним ситуации, что естественно для единицы, названной «текстовым **фактом**».

Пользователя может заинтересовать иной аспект ситуации (отличный от словарно или контекстно выделенного), в том числе предметные сущности. Например, в ситуацию вовлечены семантически сильные «предметные» участники, как то: деньги, наркотики, преступники, спонсоры и т.п. Будем обозначать «предметных» участников символами объектов (ОБ1…ОБк), но позволим им занимать позицию главного предиката при включении в состав единицы ТФ. Итак, структура

ТФ может включать и такие единицы, где онтологическим предикатом объявлен искомый объект, а его аргументами оказываются уточняющие параметры и другие сильные семантические узлы, в том числе само Событие:

ОБ=преступник(ИМЯ,КЛИЧКА,Соб=*пойман*,ВРЕМЯ).

Если пользователя интересует последовательность событий-катастроф, можно задать ТФ как цепочку из имен отрезков времени (Т1 …Тк), включающих эти события:

ТФ = Т1(СОБ1) + Т2(СОБ2) + ….+ Тк(СОБк).

Семантический анализ должен объединить построенные единицы в связную структуру ТФ, но может ограничиться простым перечислением найденных единиц (СИТ, СОБ, ОБ, Термины).

Каждый ТФ выражает в сжатом виде главную информацию текста. Это своего рода многоместный предикат, аргументы которого собраны не как заполнители словарных синтаксических валентностей предиката, а как реализованные в тексте и подтвержденные всем текстом семантические участники описываемой ситуации. В вышеприведенном примере ТФ совпадет с Событием: ***Сбили военный вертолет*** **(1Агент, 2Время, 3Место).**

Итак, собственно задачам адаптации предшествует интересная лингвистическая работа с построенным семантическим пространством текста. Способ построения крупных единиц (СИТуация, СОБытие и Текстовый Факт), имеющих значимость для «внешнего» мира, более подробно описан в одной из последних статей автора[36]. Анализ произвольных текстов с извлечением из них тех связных ситуаций, которые составляют содержательную основу текста, всегда представлялся мне задачей номер один. Из «ситуаций» как лингвистических единиц формируются более крупные единицы и структуры, которые можно считать уже информационными представлениями. Они нужны не только для гибкого механизма настройки лингвистического процессора на разные ПО, но и для решения других не просто актуальных, а «горящих», в социальном масштабе, проблем.

Всю задачу ИЛМ можно сформулировать как сложное преобразование адаптивного семантического пространства текста «навстречу» предметной области и/или индивидуальным знаниям или поисковому предписанию пользователя. Такая задача может выполняться на экспериментальной структуре связного текста, какой является структура Семантического Пространства.

Большую роль в этой теории прикладной системы выполняет Словарь с описанием возможного поведения слов именно в масштабе текста и относительно ПО. Но описание устройства семантического словаря в модели ИЛМ – отдельная задача[37].

Семантический анализ любого естественного текста (ЕТ) связан с преодолением принципиальных трудностей разного рода. Прежде всего это заведомая неполнота семантического словаря, создание которого в полном масштабе – очень трудоемкая задача. Добавим и субъективизм исходных семантических описаний слов в словаре. Обе трудности преодолимы при учете законов организации ЕТ (избыточность, синонимия и др.) как объекта анализа и благодаря выбранной методике анализа, которая позволяет работать с таким несовершенным объектом, как СемПрост текста.

8. Особенности семантического пространства как адаптивной семантической структуры

Если сравнить предполагаемые результаты семантического анализа целого текста двумя методами – ЭС («элементарных смыслов») и ЭСО («элементарных смысловых отношений») – то можно увидеть не только различия, но и сходство. В обоих случаях мы получаем своего рода семантическое пространство текста. Подход ЭС, производя декомпозицию значимых лексем на составляющие их более мелкие смыслы, как бы «размывает» границы между словами, позволяя комбинировать ЭС иначе, чем это было в анализируемом тексте. Подход ЭСО «размывает» границы между предложениями (включая и другие части текста), выравнивая структуру целого использованием только одной простейшей синтаксической конструкции. Первый подход «работает» с атомами, второй – с молекулами (целыми словами) и их связями. В обоих случаях мы имеем дело с незаконченными, промежуточными структурами, открывающими простор для использования свойств целого текста. Считаю, что оба типа структур поставляют нам ценный экспериментальный объект, позволяющий создавать и отрабатывать формальные методы анализа содержания текста.

Хотя мы остановились в практической работе на построении только промежуточной структуры, а необходимые в семантике процедуры сжатия содержания текста проследили лишь умозрительно, можно утверждать, что мы вышли в другую, отличную от классических, модель. Это несимметричная модель (анализ и синтез обслуживаются разными грамматиками).

В ней само определение «смысла» текста претерпевает изменения. Если в классических моделях СемП сохраняет все то, что было в тексте (т.е. Смысл = СемП = Текст), то в ИЛМ выполняется неравенство: Смысл короче Текста. Смысл формируется заново для каждого акта его восприятия – это функция отображения знаний и вопросов ВУ на содержание воспринимаемого текста. Он ближе к понятию Информации по Шрейдеру[38], что требует более детальной разработки, выходящей за рамки приводимых здесь беглых рассуждений.

Мы начинали в ЛМП работы по машинному переводу «через смысл», имея в виду цель, которая сначала виделась как одна для всех и единственно правильная. Эта прекрасная цель – создание СемП целого текста как однозначной, полной, правильной и логически непротиворечивой структуры – образует лишь одно из возможных «измерений» семантики текста. Желательно, чтобы хоть одна из возможных интерпретаций содержания целого текста в практической системе удовлетворяла так формулируемому идеалу.

Укажу еще два аспекта, по которым ИЛМ расходится с МСТ. Это отношение к так называемым «дефектам» естественного текста – а) неполноте и б) неоднозначности разного рода. Неполнота фиксируется нами в явном виде еще в первичной интерпретации синтаксических связей в терминах ЭСО, неоднозначность тоже отображается в той же структуре СемПрост с помощью метаотношения несовместимости двух или более частей структуры. Тема видов смысловой неполноты развита в отдельной статье[39].

В нашей модели эти «недостатки» играют конструктивную роль. Фиксация и работа с неполнотой в структурах текста моделирует работу с компонентом Знания: ведь задаваемое «сверху» знание будет всегда неполным (в том числе СемП вопроса пользователя к тексту). Структура СемПрост **динамична**. На следующих уровнях анализа неполные формулы могут заполняться единицами из других частей текста – на основании только смысловых критериев[40]. Содержательную неоднозначность локальных участков и целого СемП желательно сохранять, так как она моделирует восприятие и понимание текста, которые в норме неоднозначны.

Содержание текста по-разному сочетается со знаниями и установками воспринимающей личности (ВУ). Благодаря свойству **эластичности** СемПрост в нее могут быть добавлены (или изъяты) любые фрагменты, записанные на том же метаязыке. В описываемой абстрактной модели это даст возможность

каждой конкретной личности добавлять свой вопрос, придав большой вес искомым сущностям, и далее запускать анализ, отыскивать в тексте свои «линии интересов», или строить «свой смысл». Смысл появляется (или не появляется) при восприятии текста другой разумной системой, в том числе человеком.

И еще одно соображение хочется добавить. На мой взгляд, полный анализ текста должен сопровождаться **оценкой** всех его качеств. Это и дефекты исходного ЕТ: сколько в нем синтаксически неполных предложений, как часто нарушаются правила семантической согласованности, сколько фраз не имеют логического продолжения и т.д. Важны и положительные свойства текста (сколько и какие темы затронуты, как выделяется содержательный центр и другие). Плохое информационное качество конкретных ЕТ может дать интересный научный и даже практический выход: логическим следствием (пока ещё далеким) анализа текстов явится возможность формальной оценки информационных свойств ЕТ, в частности, не только меры правильности, но и степени связности текста, а также новизны и ценности информации, заключенной в каждом исходном тексте[41].

Работа над уточнением текстового семантического метаязыка ЭСО (основа которого описана в разных статьях автора, начиная с автореферата канд. дисс. 1968 года) имеет скорее теоретический характер, но относится не к теории устройства языка (как в ТСТ), а к теории автоматического «понимания» и восприятия текста, к прикладной лингвистике. Свойства метаязыка ЭСО (синтагматика, парадигматика и логические свойства отношений) задают ту «метрику», которая управляет возможными преобразованиями единиц на структуре СемПрост. Считаю, что уже структура СемПрост (как первая реальная в составе ИЛМ промежуточная семантическая структура целого текста) дает основания осуществлять важнейшую функцию АПТ – сжатие и компактное представление содержания текстов.

Заключение

Проследить свою собственную научно-практическую линию в составе ЛМП и в последующих работах я считала небесполезным потому, что она сложилась на пересечении нескольких отечественных направлений и/или школ, связанных с семантикой: теории элементарных смыслов, теории «Смысл↔Текст», лексической семантики и информатики. Начиная с задачи машинного перевода как научной и прикладной дисцип-

лины, эта линия выводит к комплексной проблеме, названной «искусственный интеллект», в ее гуманитарной части. К сожалению, реализовать на компьютере из семантических задач удалось немногое. Не ищу оправданий в разных внешних обстоятельствах, но одну из причин указать даже приятно – это слишком быстрый прогресс вычислительной техники. Ведь в ходе работ нам пришлось перепробовать все поколения: от ЭЦВМ, больших, миди- и мини-ЭВМ до современных ПК и ноутбуков. И здесь проблема стыковки разных систем стояла очень остро: из-за разных видов несовместимости и «непереносимости», в том числе технической, не раз приходилось бросать сделанное и начинать почти «с чистого листа». Надо сказать, что это обстоятельство тоже способствовало поиску более естественного и простого способа анализа текста, диктуемого больше материалом текстовых реалий, чем задаваемыми «сверху» законами построения только правильных структур.

Многомерная по своей сути структура текста, включающая возможные расширения и сжатия, интерпретации и выводы, обобщения и специализацию, может быть представлена как плоский вербальный объект, как «псевдотекст», записанный на некотором едином метаязыке. Этот объект, названный Семантическим пространством текста, позволяет формулировать многие свойственные человеку семантические операции гораздо проще и естественней, чем они выглядели бы в терминах строгих лингвистических правил, работающих от предложения к предложению. Например, применение аксиомы «Одно и то же лицо не может в один и тот же отрезок времени находиться в нескольких разных местах» позволит отбросить как неверные множество альтернативных формул, построенных в локальном контексте разных предложений как синтаксически правильные варианты. Конечно, одновременно должно работать и множество других правил (доказательства тождества самих «объектов» в разных упоминаниях по тексту и т.д.). Но это как раз лишний стимул подумать об иной организации собственно семантических процессоров, моделирующих понимание целого текста.

Переход от последовательной организации модулей семантического анализа к параллельной работе многих разнородных проверок в одном семантическом пространстве поможет освободить лингвистов от планирования только сверху «правильных» результатов понимания текста на каждом уровне. Ведь лингвисту приходится еще в словаре закладывать все возможности поведения слова во всех контекстах. Может быть, лучше упростить описания лексем в словаре в их синтаксиче-

ски правильном поведении и перенести центр внимания на учет свойств самого текста, полагаясь на правила «здравого смысла» и вычислительные возможности компьютера? Этот вопрос остается пока без ответа, так как он требует проведения многих экспериментальных работ.

В заключение я хочу выразить искреннюю благодарность С.И. Гиндину за его поддержку «целотекстных» и Domain-ориентированных идей автора в течение многих лет совместной работы – в ЛМП МГПИИЯ и далее на кафедре ТиПЛ РГГУ, а также за помощь в работе над моими заметками в данном разделе Вестника РГГУ.

Примечания

[1] *Perry J.W., Kent A., Berry M.M.* Machine literature searching. N.Y.: Interscience Publishers, 1956.

[2] *Жолковский А.К., Леонтьева Н.Н., Мартемьянов Ю.С.* О принципиальном использовании смысла при машинном переводе // Сб. Машинный перевод, ИТМ и ВТ АН СССР, 1961. Здесь и далее полные библ. данные ко всем упоминаемым в тексте работам автора до 2006 г. включительно, а также дополнительные сведения даются по статье: *Гиндин С.И., Семенова Н.Г.* Материалы к библиографическому указателю печатных работ Н.Н.Леонтьевой // Вестник РГГУ. 2007. №8, с. 215-235 (серия МЛЖ № 9/2).

[3] Та же статья, в кн.: *Мартемьянов Ю.С.* Логика ситуаций. Строение текста. Терминологичность слов. См. *Гиндин, Семенова.* Указ. соч.

[4] Машинный перевод: Сборник статей; перев. с англ. / Под ред., предисл. П.С. Кузнецова. М.: Изд. иностранной литературы, 1957. В сборнике статья: В. Ингве «Синтаксис и проблема многозначности» и статья Л. и А. Вундхейлеров «Некоторые понятия логики в применении к синтаксису».

[5] Автоматический перевод: Сборник статей: перев. с англ., итал., нем. и франц./ Под ред., предисл. О.С. Кулагиной, И.А. Мельчука. М., 1971. 368 с.

[6] *Мельчук И.А.* Автоматический синтаксический анализ. Т.1. Общие принципы. Внутрисегментный синтаксический анализ. Новосибирск. 1964. 359 с.

[7] *Иорданская Л.Н.* Автоматический синтаксический анализ. Т.2. Межсегментный синтаксический анализ. Новосибирск. 1967. 232 с.

[8] Историю МП см. в кн.: *Кулагина О.С.* Исследования по машинному переводу. М.: Наука, 1979. 320 с.

[9] Машинный перевод и прикладная лингвистика. Вып. 8, М.: Изд-во МГПИИЯ, 1964. 252 с.

[10] *Жолковский А.К., Мельчук И.А.* О семантическом синтезе // Проблемы кибернетики. Вып. 19. М.: Наука, 1967. С. 177-238.

[11] *Мельчук И.А.* Опыт теории лингвистических моделей «Смысл ↔ Текст». Семантика, синтаксис. М.: Наука, 1974.

[12] ТКС – толково-комбинаторный словарь. Авторы: Апресян Ю.Д., Мельчук

И.А. и др. Серия публикаций ПГЭПЛ, Ин-т русского языка АН СССР, 1976-1996. ТКС описан также И.А. Мельчуком в его книге «Опыт теории лингвистических моделей Смысл ↔ Текст». См. сноску 11.

13 *Жолковский А.К., Щеглов Ю.К.* Работы по поэтике выразительности: Инварианты – Тема – Приемы – Текст. М.: Прогресс. 1996. – 344 с. (Предисл. М.Л. Гаспарова).

14 *Мартемьянов Ю.С., Дорофеев Г.С.* Серия работ по семантике связного текста, см. в кн.: *Мартемьянов Ю.С.* Логика ситуаций. Строение текста. Терминологичность слов. М.: Языки слав. культуры, 2004.

15 *Гиндин С.И.* Внутренняя организация текста. Элементы теории и семантический анализ. Дисс. ... канд. филол. наук. М., 1971. 390 с.

16 *Шаляпина З.М.* Англо-русский многоаспектный автоматический словарь (АРМАС) // МП и ПЛ, 1974, вып.17. С. 7-116.

17 *Uchida Hirochi, Zhu Meiying,* Tarcisio Della Senta. A Gift for a Millenium. http://www.unl.ias.unu.edu.

18 *Леонтьева Н.Н., Никитина С.Е.* Смысловые отношения, передаваемые русскими предлогами. См. Гиндин-Семенова, 1969:3.

19 *Леонтьева Н.Н.* Об одном способе представления смысла текста. См. в: *Гиндин, Семенова.* Указ. соч. (1967:2).

20 *Fillmore, Charles J.* The case for case // Universals in Linguistics Theory. Austin, Texas, 1968.

21 Статьи *Леонтьева Н.Н., Никогосов С.Л.* Контуры анализа в системе французско-русского машинного перевода, Система ФРАП как информационная система, *Леонтьева Н.Н.* Система французско-русского автоматического перевода (ФРАП): лингвистические решения, состав, реализация см. в: *Гиндин, Семенова.* Указ. соч. (1977:1, 1982:7, 1987:3).

22 Моделирование языковой деятельности в интеллектуальных системах // Под ред. Кибрика А.Е., Нариньяни А.С. М.: Наука, 1987. 280 с.

23 См. два выпуска «Обзорной информации» ВЦП: 1) Синтаксический компонент в системах машинного перевода, 2) Семантический компонент в системах автоматического понимания текстов в: *Гиндин, Семенова.* Указ. соч. (1981:3, 1982:5).

24 *Леонтьева Н.Н.* Информационная модель системы автоматического перевода. *Leontyeva, Nina.* Stages of Information Analysis of Natural Language Texts. См. в: *Гиндин, Семенова.* Указ. соч. (1985:1; 1987:5,6).

25 *Гиндин С.И., Леонтьева Н.Н.* Задачи и общее строение системы автоматического индексирования с использованием информационного языка словосочетаний см. в: *Гиндин, Семенова.* Указ. соч. (1975:1).

26 Статьи *Леонтьева Н.Н., Эрастов К.О.* Автоматизация обработки текстовой части экономической информации для информационной системы, *Леонтьева Н.Н., Мартемьянов Ю.С., Розенцвейг В.Ю.* О выявлении и представлении смысловой структуры текстов экономических документов см. в: *Гиндин, Семенова.* Указ. соч. 1(967:1; 1971:2).

27 *Леонтьева Н.Н., Урысон Е.В.* Алгоритм построения информационной записи для текста (1 этап). См. *Гиндин, Семенова.* Указ. соч. (1973:1,2).

28 *Леонтьева Н.Н., Вишнякова С.М.* Опыт автоматического индексирования со смысловым сжатием. См. *Гиндин, Семенова.* Указ. соч. 1(977:2).

29 *Леонтьева Н.Н., Волковыская Е.В., Копылова О.Т., Молчанова Т.В., Штернова О.А.* Словарь энциклопедических функций и его роль в авто-

матическом индексировании. См. *Гиндин, Семенова.* Указ. соч. (1978:2).

30 *Леонтьева Н. Н.* Автоматическое понимание текста: системы, модели, ресурсы. См. *Гиндин, Семенова.* Указ. соч. (2006:1).

31 *Никитина С.Е.* Тезаурус по теоретической и прикладной лингвистике. М.: Наука, 1978. 374 с.

32 *Леонтьева Н.Н.* ПОЛИТЕКСТ: информационный анализ политических текстов. См. *Гиндин, Семенова.* Указ. соч. (1995:1).

33 *Леонтьева Н.Н.* К теории автоматического понимания текстов. Часть 1. Моделирование системы «мягкого понимания» текста: информационно-лингвистическая модель. См. *Гиндин, Семенова.* Указ. соч. (2000:1).

34 *Кобзарева Т.Ю.* Принципы сегментационного анализа русского предложения // Московский лингвистический журнал. М.: Изд-во РГГУ, 2004. Т. 8. №1. С.31-80.

35 *Ермаков М.В.* К выявлению лексического ядра лингвистической ситуации (на материале текстов криминальных сводок). См. настоящий выпуск.

36 *Леонтьева Н.Н.* Постсемантический анализ текста: промежуточные структуры //Динамические модели: Слово. Предложение. Текст: Сб. статей к юбилею Е.В.Падучевой. М.: Языки слав. культуры. М., 2008. С. 526-546.

37 *Леонтьева Н.Н.* Русский общесемантический словарь (РОСС): структура, наполнение. См. *Гиндин, Семенова.* Указ. соч. (1997:1).

38 *Шрейдер Ю.А.* О семантических аспектах теории информации// Информация и кибернетика. М., Наука, 1976. С. 15-47.

39 *Леонтьева Н.Н.* Смысловая неполнота или неграмотность? Взгляд прикладного лингвиста // Slavica Helsingiensia 35. С любовью к слову: Festschrift for Arto Mustajoki. Helsinki 2008. P. 144-167.

40 *Леонтьева Н.Н.* Семантический анализ и смысловая неполнота текста. См. *Гиндин, Семенова.* Указ. соч. (1968:1).

41 *Леонтьева Н.Н., Никогосов С.Л.* Система ФРАП и проблема оценки качества автоматического перевода. См. *Гиндин, Семенова.* Указ. соч. (1980:1).