ВЫСШЕЕ ПРОФЕССИОНАЛЬНОЕ ОБРАЗОВАНИЕ

Н. Н. ЛЕОНТЬЕВА

АВТОМАТИЧЕСКОЕ ПОНИМАНИЕ ТЕКСТОВ

СИСТЕМЫ, МОДЕЛИ, РЕСУРСЫ

Для студентов
лингвистических факультетов вузов

Москва
ACADEMIA
2006

УДК 800(075.8)
ББК 81.1я73
Л478

Рецензенты:
доктор филологических наук, профессор, зав. кафедрой лингвистической семантики Московского государственного лингвистического университета *Б. Ю. Городецкий*;
доктор филологических наук, профессор, главный научный сотрудник Института русского языка им. В.В.Виноградова РАН *В. М. Андрющенко*

Леонтьева Н. Н.

Л478 Автоматическое понимание текстов: системы, модели, ресурсы: учеб. пособие для студ. лингв. фак. вузов / Нина Николаевна Леонтьева. — М.: Издательский центр «Академия», 2006. — 304 с.

ISBN 5-7695-1842-1

Учебное пособие обобщает опыт создания отечественных и зарубежных систем, реализующих автоматическое понимание текстов. Эти сложные «интеллектуальные» системы выделяются из множества систем, в которых просто используется автоматическая обработка текста, поскольку автора интересует именно качественный аспект понимания. Рассмотрены те компоненты процесса АПТ, которые могут быть заданы в вербальном виде. В основе пособия — идея «мягкого» понимания текста; представлена экспериментальная лингвистическая система ПОЛИТЕКСТ, осуществляющая гибкое соединение лингвистических и предметных знаний.

Для студентов лингвистических факультетов вузов. Может быть рекомендовано для тех, кто интересуется искусственным интеллектом, структурной и прикладной лингвистикой, информатикой.

УДК 800(075.8)
ББК 81.1я73

ISBN 5-7695-1842-1

ПРЕДИСЛОВИЕ

В данном учебном пособии прослежен лингвистический аспект учебной дисциплины «Автоматическая обработка текстов» (АОТ). На примере некоторых представительных систем и моделей рассматриваются компоненты, из которых складывается полный цикл процесса **автоматического понимания текста** (АПТ), и лингвистические ресурсы, необходимые для его компьютерной реализации. На фоне беглого изложения разных подходов к решению лингвистических проблем построения систем АПТ даются развернутые иллюстрации из авторских работ. Рассматриваются отдельные трудные участки процесса АПТ. Такой жанр обсуждения с читателем трудностей, а не сообщения готовых истин вполне оправдан, так как в проблемной области под названием АПТ еще слишком велик разброс мнений по ее ключевым проблемам: так, не сформировалось единого взгляда на природу метаязыка (интерлингвы), на котором желательно представлять содержание любого текста; не определен статус баз знаний; нет согласия в том, что считать собственно семантическим компонентом понимания и где границы его компетенции, и т.д.

В книге представлен авторский взгляд на состав и роль **семантического компонента** как наиболее содержательного участка компьютерного понимания, определяющего функции всех остальных. Опыт преподавания различных прикладных дисциплин на отделениях структурной и прикладной лингвистики (курс «Информатика» на факультете совершенствования переводчиков МГПИИЯ, основные и спецкурсы в РГГУ и МГУ по прикладной семантике, информационному анализу текста, системам генерации текстов, словарям для систем АПТ и др.) убедил автора в том, что студентов интересуют не столько уже реализованные системы (индексирования, поиска, морфологического анализа, коррекции ошибок и т.п.), сколько вопросы о том, как повысить уровень понимания текстов в прикладных системах. При тестировании программ синтаксического анализа после ввода фраз типа *Дети пошли в школу* студенты быстро переходят к очень сложным примерам, нащупывая слабые места парсинга (синтаксического анализатора) или системы машинного перевода. Им мало интересен вывод *Петя поглощает яблоки* → (значит) *Мальчик ест фрукты*, а ведь даже такие простые выводы требуют введения сложных семантических данных в систему АПТ.

Однако можно задать множество тонкостей и деталей своей (лингвистической или переводческой) науке, но не владеть искусством синтеза этих знаний в практических компьютерных задачах из-за отсутствия цельного взгляда на комплекс проблем, связанных с содержательной обработкой «живых» текстов и массивов. Поэтому пособие адресовано в основном тем из студентов, лингвистов и программистов, кто ищет новые интересные решения трудных семантических проблем, а тем более возможности их реализации. Не последняя задача данного учебного пособия — вызвать интерес молодых специалистов к гибкому включению лингвистической теории в проектирование компьютерных интеллектуальных систем.

Структура книги. В **главе 1** обосновывается теоретическая модель, которая легла в основу учебной экспериментальной системы ПОЛИТЕКСТ.

Стержень, на котором держатся излагаемые частные решения, — это идея «**мягкого» понимания** текста. Коротко, она выражается в том, что один и тот же текст допускает разные результаты понимания в зависимости от разных условий и составляющих процесса понимания. Лингвистический характер системы и всех ее механизмов состоит в том, что мы всегда имеем дело со сравнением и преобразованием **текстов** и **текстовых структур**.

Совокупность «текстов» (в широком смысле, включая тезаурусы и списки, текст либо структуру вопроса и лексикон баз данных), вовлеченных в сеанс получения информации из заданного источника, образует информационное пространство (ИнфПрост) текста. Минимальный состав ИнфПрост — сам анализируемый текст и «встречный» текст (например, текст вопроса), в единицах которого должны пониматься исходный текст и строиться информация.

Каждый уровень понимания имеет свою сферу действия и вычленяет специфические для данного уровня единицы.

Глава 2 посвящена краткому описанию традиционных систем автоматического, или машинного, перевода (МП). Это первая наиболее полная разновидность АПТ-систем. Принятое нами расширение стандартной переводческой модели стало основой системы ПОЛИТЕКСТ и схемы мягкого понимания.

Главы 3, 4, 5 и 6 освещают конкретные аспекты разработки экспериментальной лингвистической системы ПОЛИТЕКСТ, которая ответственна за гибкое соединение лингвистических и предметных знаний. **Глава 3** описывает специфику анализируемых массивов текстов и работу первой подсистемы — графематического анализа. В **главе 4** рассматриваются способы морфологического анализа, решения (в том числе нестандартные) задач лексического и морфологического этапов анализа и их представлений, а также те семантические проблемы, которые возникают

уже на двух ранних стадиях. **Глава 5** освещает проблематику синтаксического анализа (СинАн) предложений. В составе предложенной автором информационно-лингвистической модели (ИЛМ) синтаксис рассматривается как опора для понимания и поэтому может быть неполным. Объясняется механизм взаимодействия синтаксической и семантической структур, при котором построенные формально синтаксические связи интерпретируются в первичном семантическом представлении (СемП), строятся семантические узлы, после чего можно вернуться к синтаксическому представлению (СинП) и достроить единицы СинП, уточнив их как члены предложения, аттестованные семантически (как обстоятельства места, времени и т. п.). **Глава 6** подробно разбирает состав и функционирование семантического компонента системы в составе ИЛМ. Это только локальный (в пределах каждого отдельного предложения) анализ, но он вводит свой метаязык, с помощью которого строятся все дальнейшие семантические представления.

Семантический компонент — центральная часть информационно-лингвистической модели, объясняющей мягкое автоматическое понимание текста. Он обеспечивает функции многомерного, неоднозначного, неполного, выборочного и других видов естественного понимания. Его основное назначение — справляться с разноязычием, которое может быть естественно-языковым (например, английский текст — русский реципиент), профессиональным, проблемно-ориентированным (поскольку каждая предметная область или задача вводит свой профессиональный язык) и ценностным (у каждого реципиента свои информационные установки и ценности, «в пользу» которых может строиться СемП).

Последовательность всех локальных интерпретаций единиц текста в виде **семантического пространства** (СемПрост) текста — первая реальная, строящаяся программно, а не только теоретически целотекстная структура.

Описанный в **главе 7** процесс создания **глобальной структуры** текста сводится с технической точки зрения к устранению свойств неидеальности СемПрост текста. Это устранение избыточности, уточнение отношений иерархии, устранение единиц, получивших в ходе анализа малый информационный вес, и др. Основная содержательная операция глобального анализа — синтез новых единиц типа СИТ (**ситуация**). Высшей единицей, представляющей текст во внешней среде, предложено считать **текстовый факт** (ТФ). Это и лингвистическая единица, и кандидат на включение в базу знаний определенной предметной области (ПО), т. е. объект, традиционно относящийся к экстралингвистическим единицам.

Лишь на глобальной структуре могут проводиться важнейшие процессы сравнения содержания разных текстов, приводящие к построению «текста информации». Это структура, получаемая в

результате «вычитания» СемП вопроса из СемП текста, из которой можно сгенерировать текст ответа пользователю.

В **главе 8** показано несколько способов структурирования специальных знаний, характеризующих ту или иную предметную область. В основном это массивы предметной, энциклопедической информации, собранные в базы данных (БД). Это и словари терминов данной ПО, и БД, в которые включены имена и описания разных спецобъектов в их иерархических связях. Подробно описан Тезаурус общественно-политической терминологии (РуТез), являющийся главным компонентом действующей системы РОССИЯ как инструмент индексирования, рубрицирования и информационного поиска.

Все описанные в главе 8 системы иллюстрируют не только способы представления специальных знаний, но и разные приемы **смыслового сжатия текста**.

Глава 9 посвящена описанию нескольких информационных процессов, родственных процедуре автоматического сжатия текстов. Это системы индексирования, рубрицирования, аннотирования и реферирования; к получаемому результату обычно добавляют элемент *квази*: квазианнотация, квазиреферат и т.д. Основную роль в этих процессах играет морфологический, статистический и терминологический анализ. Эти работы очень важны, так как они имеют дело с естественным входным материалом (действительно значительно сжимая его), с естественным пользователем, а главное, они надолго определили методику работы с массивами текстов. Как и системы машинного перевода, они проложили путь следующим более интеллектуальным системам АПТ (потенциал которых, однако, еще не реализовался), это системы *Information Extraction* (*IE*), *text mining*, *data mining*, *knowledge discovery*, *knowledge aquisition* и др. В общем виде их задача состоит в извлечении частичных знаний из больших массивов текстов, т.е. в обнаружении таких фрагментов текста, которые отвечают заданной информационной установке и могут быть помещены в формат баз данных.

Глава 10 кратко описывает системы генерации текстов (СГТ) и содержит обзор проблем, встающих при синтезе информации из разных типов баз данных и представлении результатов в виде естественных текстов (ЕТ). СГТ обеспечивают автоматическое порождение связных текстов на естественном языке (ЕЯ). Основой для работы СГТ могут служить разнообразные семантические и концептуальные структуры, в частности базы данных и знаний, из которых пользователь хочет извлечь интересующую его информацию уже в словесном виде.

Те параметры текста, которые нужно учитывать при генерации, обостряют требования к семантическому анализу и формированию его результатов.

По сути дела, СГТ — это вторая часть того подхода, который принят в системах АПТ типа «перевод-реферат на основе базы знаний».

Глава 11 завершает рассмотрение систем и методов АПТ авторским эскизом такой интегральной информационно-переводческой системы, работа которой состоит в постоянном наполнении **Базы текстовых фактов** (БТФ). «Текстовой» она названа не только потому, что создается при анализе текстов и включает те знания, которые несет конкретный текст, но и потому, что значения ее полей могут быть заполнены свободным текстовым материалом, не скованным заранее заданными форматами в основном количественного характера, как в стандартных БД. БТФ собирает информацию из текстов общего пользования и гуманитарного характера. В качестве примера были проанализированы тексты СМИ, дана иллюстрация возможных результирующих записей в такой БТФ. Архитектура БТФ позволяет использовать ее в разных прикладных задачах: реферирования, перевода, информационного поиска и обобщения данных, синтеза на ее основе других частных баз данных, генерации новых текстов и т. п. В задаче построения БТФ и производных от нее «продуктов» и подсистем должны быть использованы все наработанные к настоящему моменту и охарактеризованные в данном пособии лингвистические ресурсы.

Глава 12 дает представление о комплексе словарей, обслуживающих систему АПТ полного состава (от ЕТ до получения Информации адресатом). Последовательность процедур анализа должна обслуживаться комплексом соответственно ориентированных словарей, каждый из которых имеет дело со специфическими для данного уровня единицами и информацией к ним. Подробно описан основной инструмент семантического анализа текста в системе ПОЛИТЕКСТ — словарь РУСЛАН, поскольку состав полей этого словаря близок к универсальному. Он содержит исчерпывающую лингвистическую информацию, сведения об энциклопедических и информационно-тезаурусных связях слов, а также данные, позволяющие строить единицы типа Ситуация и настраивать словарь на разные задачи. Описание слов в словаре ведется «сверху вниз», в соответствии с уровнями лингвистического анализа. Основная направленность разрабатываемой версии РУСЛАНа — обеспечить построение Базы текстовых фактов для заданного массива русских текстов. Продолжением этого общеязыкового словаря являются спецсловари того типа, который описан в главе 8.

Глава 13 называет те лингвистические ресурсы, на которые может опираться любая система АПТ (*reusable resources*), к их числу относятся массивы одно- и двуязычных общих словарей, терминологических словарей и тезаурусов. Но главным ресурсом считаются большие корпуса собственно текстов, их изучением зани-

мается новая дисциплина «Корпусная лингвистика». Хотя она и не имеет непосредственного отношения к системам АПТ, но пользуется многими ее плодами, например, создавая специальные производные корпуса аннотированных текстов.

О литературе. В конце каждой главы дана рекомендуемая литература по теме. Не все источники имеют одинаковую ценность, так как в АПТ как дисциплине рано ставить хотя бы временную точку. Скорее, это перечень источников, на основании которых сложились авторские обобщения. Работы не делятся на обязательные и факультативные, в списках к разным главам возможны повторы. Для учебных целей достаточно выборочного знакомства с двумя-тремя публикациями по каждой теме. В качестве рекомендуемой литературы приводятся и работы, которые были выполнены в ранний период довольно идеалистических представлений о возможностях компьютеров, но и в старых работах можно найти иногда добротные и полезные решения по лингвистическому обеспечению систем АПТ. Частые ссылки на работы автора объясняются жанром (обобщение спецкурсов) и доступностью собственного материала, иллюстрирующего идею книги.

Вопросы по всем главам, необходимые в жанре пособия, расположены перед приложением.

* * *

Хочу поблагодарить заведующего кафедрой теоретической и прикладной лингвистики РГГУ С. И. Гиндина за идею изложить основные положения спецкурсов в виде отдельного пособия. Ряд семантических положений докладывался и обсуждался на семинарах под руководством А. И. Новикова в Институте языкознания РАН. Книга была прочитана рецензентами В. М. Андрющенко и Б. В. Городецким, а также С. Е. Никитиной, заострившей ряд интересных теоретических вопросов, и Н. В. Перцовым, высказавшим замечания по главе 5. Всем им я признательна за поддержку жанра пособия (авторское обобщение подходов к созданию систем АПТ) и полезные советы и замечания. Выражаю благодарность Л. Н. Иорданской, Р. Киттреджу, С. Ниренбургу и А. Я. Шайкевичу за помощь с современной литературой и материалами конференций, которые позволили следить за уровнем разработок по теме систем АПТ. Моим коллегам по инициированной Т. Н. Юдиной работе над системой ПОЛИТЕКСТ (программистам Ж. Г. Аношкиной, А. В. Сокирко и др., лексикографам М. Г. Шаталовой, Е. М. Сморгуновой, С. Ю. Семеновой, А. С. Паниной, Е. В. Горелик, равно как всем участникам проектов прошлых лет — они названы в соответствующих разделах книги) особое спасибо за преданность идее семантического анализа и вклад в конкретные результаты. Также я признательна руководству НИВЦ МГУ им. М. В. Ломоносова и сво-

им коллегам за поддержание обстановки, благоприятной для научных работ и полезных контактов со студентами.

Студенты РГГУ тоже внесли лепту в развитие системы своим участием в дискуссиях и вводе словарных статей в БД Русский общесемантический словарь (РОСС), а также рядом курсовых и дипломных работ. В 2000—2001 гг. группой выпускников РГГУ была создана на основе проекта ПОЛИТЕКСТ пробная версия системы МП с русского языка на английский ДИАЛИНГ, а затем и ее версия в сети Интернет (www.aot.ru). Она была выполнена на современном программистском уровне с добавлением этапа собственно перевода, с рядом изменений, с упрощением синтаксиса и др. В нем ценен семантический компонент, который берет на себя часть проблем, не решенных синтаксическим анализом, хотя система и не была доведена до окончательного вида по экономическим причинам.

Без поддержки неформального коллектива рядом грантов работа не могла быть проделана. Так, работы по проекту ПОЛИТЕКСТ велись при поддержке грантов Фонда Макартуров (до 1996 г.), а также Российского фонда фундаментальных исследований (РФФИ: 97-06-80093 и 99-06-80296а «Исследование информационных свойств естественного текста методом построения лингвистических структур»). Работы по созданию компьютерной базы РОСС поддерживались также Фондом «Культурная инициатива» (в 1995 г.) и Российским гуманитарным научным фондом (РГНФ: 96-03-12103в) до конца 1999 г. В 2001—2004 гг. работы по ведению и развитию автономной словарной базы, названной «РУСЛАН-1», поддерживал грант РГНФ (01-04-16252а). Следующий грант РГНФ 04-04-00185а выделен коллективу на поддержание работ по развитию и формализации метаязыка семантических и концептуальных отношений.

Автоматическая обработка или понимание текста?

В предлагаемом учебном пособии сделана попытка обобщить опыт создания отечественных и зарубежных систем, реализующих **автоматическое понимание текста**. К ним относятся системы машинного перевода, системы автоматического индексирования, системы информационного анализа массивов официальных документов и текстов СМИ, фактографические системы, системы общения на естественном языке с базами данных и знаний и другие сложные интеллектуальные системы. Они выделяются из множества систем, в которых просто используется автоматическая обработка текста, включающая техническое сжатие текста, сортировку слов по частоте, длине и т. п., любой статистический анализ, исправление грамматических ошибок и другие частные задачи, а также различные исследовательские приемы работы с текстом, выполняемые на компьютере. Нас интересует не количественный, а качественный аспект понимания. В системах АПТ действительно моделируются некоторые функции человеческого понимания, а общение с ЭВМ предполагает использование естественного языка на входе и/или на выходе работы системы. Если некая система АПТ реализовала лишь один такт понимания, она должна хотя бы в модели объяснять весь цикл процесса, в который встраивается этот такт АПТ. В отличие от многих других систем АОТ системы АПТ обладают максимальным набором лингвистических компонентов — это **полные системы**.

В центре внимания — лингвистический аспект

Подводя итоги полувековому опыту исследований и работ по автоматической обработке текста, приходится признать, что технологии существенно опередили содержательный аспект: наработано очень много отдельных приемов и методов обработки текстов без объяснения их функционирования в составе целой **системы, понимающей текст**. Конечно, естественные тексты слабо поддаются формализации, и поэтому во многих сложившихся и действующих подходах преобладают эмпирические решения.

В пособии не ставилась задача рассмотреть все идеи и решения, которые были воплощены в какую-либо систему АПТ, но по возможности охвачено все разнообразие типов систем, включая системы анализа и системы генерации текстов последних лет. Пест-

рую и неравномерную картину лингвистического обеспечения в разных системах (с учетом тенденций и провозглашаемых намерений авторов систем АПТ) мы сочли полезным «дотянуть» до целостной, связной модели, воссоздав некую идеальную систему общения человека с компьютером для получения важной для пользователя информации из массива естественных текстов.

Лингвистическому обеспечению системы АПТ приходится брать на себя решение тех задач, которые поставлены теоретической лингвистикой последнего времени. Как включить в систему такие неформальные составляющие, как действительность, автор текста, адресат текста? Лингвистическое решение состоит в том, что к рассмотрению принимаются лишь такие компоненты процесса АПТ, которые могут быть заданы в **вербальном виде**. Это означает, что они могут быть учтены системой, если они заданы в виде текстов или соотносимых с ними структур, поскольку лингвистическое обеспечение системы оперирует только с текстовыми объектами.

Начиная с семантического компонента системы мы вступаем в область мало изученного или неустоявшегося: ведь даже термин «семантическое представление» (СемП), который с легкой руки И. А. Мельчука вошел в обиход компьютерной лингвистики, понимается создателями систем АПТ неоднозначно, а такими понятиями, как представление знаний, теоретическая лингвистика вовсе не занималась, их определение, а также способы построения обычно отдаются на откуп когнитологам и специалистам в определенных узких областях знаний. Между тем в такой важной и актуальной сфере, как автоматическое **извлечение знаний** из массивов естественных текстов, трудно ждать успеха без серьезной лингвистической основы. Теория или модель должны учитывать все реалии процесса АПТ — от особенностей поступающего в систему массива текстов до представления результата понимания, выдаваемого пользователю.

О модели

В книге предложен один из вариантов цельной модели АПТ. Это модель «мягкого» понимания текста. Концепция АПТ, являющаяся основой предлагаемого учебного пособия, отрабатывалась в процессе последовательных работ над разными типами систем, руководителем и непосредственным участником которых был автор. Как результат обобщения или как теоретическое обоснование разных типов прикладных систем предложена абстрактная модель, названная **информационно-лингвистической моделью**, в рамках которой прослеживаются по шагам все звенья автоматического понимания ЕТ. Начиная с семантического компонента и далее проводится авторский взгляд на состав системы АПТ, намечают-

ся способы построения таких компонентов модели, как Информация и Смысл текста. Суть концепции состоит в объяснении (применительно к системе АПТ) такого естественного феномена, как *неоднозначное восприятие текста*, состоящего в том, что разные пользователи извлекают свою индивидуальную информацию и свой индивидуальный смысл из одного и того же текста. Для этого требуется соединить *лингвистические* механизмы понимания, стремящиеся к точности и сохраняющие эквивалентность при всех преобразованиях, с *информационными*, моделирующими устранение сведений, не нужных пользователю.

В рамках принятой модели процесс анализа заканчивается построением множественной, или многомерной, структуры, в которой представлены разные возможные прочтения заданного текста. Неоднозначность (в широком смысле) мы считаем не досадной помехой в системе АПТ, но скорее конструктивным фактором, помогающим моделировать построение разных индивидуальных интерпретаций текста, разных информаций и индивидуальных смыслов. Такая трактовка понимания противопоставлена жесткому соответствию «один текст — один смысл», где смыслом объявляется семантическое представление текста, т. е. одна, хотя и сложная, формула. В нашей модели понимания СемП, как бы его ни определять, — это еще не Смысл.

Согласно предложенной модели компьютерного понимания именно во взаимодействии лингвистических уровней проявляется **механизм смыслообразования**; эта проблема представляется автору необходимой составляющей также и теории лингвистической семантики.

О проекте ПОЛИТЕКСТ

Одна из установок пособия — подавать материал не как конгломерат разных возможных приемов анализа, а как последовательность или более сложное объединение частей (компонентов), выполняющих функционально различные задачи в составе единого механизма, имеющего заданную цель. Для этого рассматривается учебная экспериментальная система, которой мы дали имя ПОЛИТЕКСТ, и большинство примеров приводится из этой не до конца реализованной системы автоматического понимания текстов на русском языке. В ее рамках иллюстрируются основные положения информационно-лингвистической модели.

ПОЛИТЕКСТ — это система?

Слово «система» имеет в контексте компьютерной лингвистики два значения. В первом значении это конкретный отлаженный и работающий на ЭВМ комплекс программ, выполняющий оп-

ределенную задачу, принимающий на входе данные в определенном формате, и т.д. Как правило, такой комплекс имеет полное узаконенное технологическое обеспечение, начиная от способа получения входных данных и кончая установленным кругом пользователей, с учетом требований которых этот комплекс создавался. Такой комплекс всегда имеет сопровождение, т.е. группу работников, обеспечивающих его бесперебойное тестирование и функционирование. В группу сопровождения обычно входят и авторы разработки, которые, во-первых, ответственны за концептуальную сторону системы, а во-вторых, получают полезные сведения о результатах работы, о том, как ее оценивают пользователи и т.д., — все эти сведения могут быть учтены для улучшения данной или при разработке следующей версии системы. «Система» в таком понимании имеет обычно собственное имя, чему мы знаем множество примеров (системы машинного перевода СИСТРАН, АРИЭЛЬ, ЭТАП, ФРАП, ЯРАП, ПРОМТ, а также ИПС: СКОБКИ, БИТ и др.).

Во втором значении это слово не имеет конкретного денотата, а обозначает нечто целое, т.е. организованную по определенным законам совокупность, сложно организованный абстрактный комплекс, а потому требует указания при нем конкретной сущности, ср.: *система взглядов*, *бухгалтерская система*, *система правил*, *взаиморасчетов* и т.д. В этом абстрактном смысле система не имеет имени.

Работы над проектом ПОЛИТЕКСТ велись с начала 1990-х гг. в рамках Центра информационных исследований (ЦИИ), который был учрежден в Институте США и Канады РАН (ИСКРАН). Прежде чем реализовывать систему в первом значении, т.е. как действующее устройство, мы спроектировали систему во втором значении этого слова — как систему установок и решений, позволяющих видеть проблему АПТ достаточно широко.

Выросшая из одного корня, работа расслоилась на два потока: лингвистическое и тематическое (информационное) направления в анализе массивов политических текстов. По замыслу, они должны были дополнять друг друга, а будущей системе было дано имя РОССИЯ. При переходе в НИВЦ МГУ было сохранено одно, информационное, направление, ставшее университетской информационной системой РОССИЯ (УИС РОССИЯ).

На основе лингвистической ветви проекта ПОЛИТЕКСТ в 2000—2001 гг. была создана пробная версия системы МП с русского языка на английский ДИАЛИНГ, в которой реализован полный цикл понимания с семантическим компонентом, опирающимся на Русский ОбщеСемантический Словарь и построенный по его модели английский словарь [Сокирко, 2001]. В рассыпанном виде проект ПОЛИТЕКСТ продолжает жить и развиваться в составе разных коллективов и систем.

Использование в данном пособии термина «система» в первом значении не только естественно, но и удобно по ряду соображений. Во-первых, в терминах «система», «подсистема», «модули», «процессоры», «словари», «базы данных» и т.п. современному читателю гораздо легче понимать и представлять себе строение этой будущей системы, а автору удобнее описывать ее. Во-вторых, конечной, пусть и отдаленной, целью всех работ является единая многоканальная система понимания текстов и автоматического извлечения информации из текстов. В-третьих, имя ПОЛИТЕКСТ (сложный, многослойный текст; много текстов в одном тексте) дает представление о главной презумпции используемой модели и реализованной части разработки: допущение неоднозначного понимания одного и того же текста (работа начиналась с анализа политических текстов) разными читателями/пользователями. Текст может быть понят с разной степенью подробности, с разными оценками, с разными фокусами внимания, с точки зрения разных предметных областей. Для автоматических процедур это очень сложная задача, как минимум, она требует распараллеливания работы лингвистических процессоров и обмена результатами. Но предлагая авторский взгляд на те компоненты модели, для которых нет пока хороших решений, мы считали необходимым рассуждать с позиций строящейся и в принципе реализуемой **системы в первом значении** (этим объясняется частое обозначение такого виртуального объекта словом, написанным с заглавной буквы: *Система*). И она должна уметь развиваться, чтобы обладать семантической силой, достаточной для полезного ее функционирования. Даже в периоды вынужденной безработицы (отсутствие сложных и дорогостоящих экспериментов) уровень интеллектуальности системы АПТ должен повышаться; многие гипотезы могут быть проверены теоретически и «вручную».

Место семантики

От выбранной семантической модели зависят организация всех других компонентов, цели каждого из них. Поэтому проектирование системы АПТ предлагается начинать «сверху вниз», определив сначала функции **семантического компонента**. В цепи стадий понимания текста ему принадлежит основная роль. При описании компонентов системы, предваряющих семантический анализ, внимание уделяется не столько технологическим и реализационным аспектам обработки текстовых элементов, сколько роли каждого из них в осмыслении текста: уже начиная с графематического уровня каждый выделенный элемент имеет шанс попасть в целевые (т.е. семантические и концептуальные) структуры текста.

ГЛАВА 1
ВЗГЛЯД «СВЕРХУ» НА СИСТЕМЫ АВТОМАТИЧЕСКОГО ПОНИМАНИЯ ТЕКСТА

§ 1. Прикладная и теоретическая лингвистика

Феномен человеческого понимания объясняет в первую очередь теоретическая лингвистика; ее роль возрастает, если объектом понимания являются тексты на естественном языке. Если мы хотим какой-то из процессов понимания реализовать на ЭВМ, необходимо обращение к компьютерной лингвистике.

Новая ветвь лингвистики — современная **компьютерная лингвистика** (КЛ), или **вычислительная лингвистика** (ВЛ), или **алгебраическая лингвистика** (термин, используемый в Пражской школе) — утвердилась, когда вошли в жизнь компьютеры и была осознана необходимость и возможность не только хранить, но и перерабатывать с их помощью большие массивы информации. ВЛ и КЛ включаются в более широкую дисциплину — **Прикладная лингвистика** (ПЛ) [см.: Баранов, 2001], или прикладное языкознание [см.: Бондарко, Вербицкая, Мартыненко и др., 1996]. Из всех прикладных проблем мы будем рассматривать лишь решение задач, которые ставятся перед так называемыми «интеллектуальными» системами **автоматической обработки текста**, реализующими **автоматическое понимание текста**. Если ниже мы используем термины и сокращения ПЛ, ВЛ и КЛ, то в этом суженном смысле и как синонимы. При этом для обозначения дисциплины, занимающейся разработкой систем АПТ, ВЛ [см.: Демьянков, 1985; и др.] предпочтителен как более точный и неомонимичный (ведь аббревиатурой КЛ обозначается теперь и корпусная лингвистика (см. гл. 13). Термин «вычислительная лингвистика» принят также в книге Р. Шенка [см.: Шенк, 1980].

Чтобы создать общую модель процесса автоматического понимания, нужна теория, объясняющая, на какие кванты делится процесс понимания и какие именно звенья этого процесса можно передать автомату. Как же в этой новой теории взаимодействуют теоретическая и прикладная вычислительная лингвистика?

Существенная часть понятий ВЛ совпадает с теми, которые использует современная **теоретическая лингвистика** (ТЛ) как наука о естественном языке, но многое приходится переопределять или уточнять. Ведь компьютерная лингвистика вынуждена форма-

лизовать все исходные понятия и все шаги их обработки. Наибольшие различия между ТЛ и ВЛ наблюдаются на первых, простейших уровнях анализа (так, понятие «слово» по-разному определяется в них и с формальной, и с содержательной точек зрения), а также на последних, когда появляются понятия «смысл», «информация», «знание». Архитектура и общее строение системы АПТ часто зависят от того, как определены важнейшие лингвистические объекты.

Системы АПТ развиваются быстрее, чем обеспечивающая их теория. Авторы систем АПТ слишком вольно пользуются словами *смысл*, *знание*, *семантический анализ* и другими, обозначающими высшие уровни понимания текста. Эти понятия затрудняется определить однозначно и теоретическая лингвистика, вернее, и ТЛ, и ВЛ дают много разных определений. Так, иногда разработчики могут назвать только морфологический анализ «смысловой обработкой» текста, но этому этапу АОТ нужно еще долго добираться до смысла. Понятие «смысл» вряд ли скоро получит формальный статус.

Отношение к **действительности** является еще одним источником расхождений ТЛ и ВЛ. Принятое в ТЛ определение ***значения*** как «отношения знака к действительности» неприменимо к ВЛ: в компьютер нельзя поместить никакой «фрагмент действительности». Но это определение можно адаптировать, применив к «миру текстов» (кодов, знаков, адресов, структур, вычислительных операций). Так, ***значение слова*** в системе АПТ — это либо тексты всех его словарных статей, либо часть текстовой структуры, соответствующая одному из выбранных (или вычисленных системой) вариантов словарных описаний.

Чтобы объяснить (и в дальнейшем ввести в Систему) названные понятия из области семантики, необходимо включить в модель АПТ некоторое воспринимающее устройство (ВУ), или адресата информации, того, кому адресованы знания, информация и который увидит или не увидит «смысл» в продукте, созданном системой понимания текста. То же относится и к автору текстов, вводимых в ЭВМ для обработки. Эти сведения, равно как «момент речи», «знание» и т.д., есть те самые элементы действительности, без обращения к которым трудно или невозможно объяснить многие собственно языковые и текстовые значения.

Автор текстов, адресат текстов, их интересы и цели — это новые объекты, которые вошли в науку последних десятилетий. Эти составляющие человеческого понимания уже подробно описаны в ТЛ и даже образовали отдельную дисциплину «Прагматика». КЛ/ВЛ, которая тоже постепенно осваивает эти сложные понятия, предоставляет богатую экспериментальную базу для их теоретического осмысления и введения более глубоких, чем синтаксис, уровней описания языка. Вопрос в том, как именно их можно ввести в систему. Полезные уточнения могут быть найдены в

опытах реализации цельных систем АПТ. **Больше всего сейчас недостает как раз моделей АПТ высших уровней.**

Итак, на вопрос *Что же **вычисляет** вычислительная лингвистика*? хотелось бы услышать такой ответ:

ВЛ вычисляет:

а) Информацию, которую передает текст и массив;

б) Смысл, который имеет эта Информация для данного (произвольного) реципиента;

в) Знание, которое адресат/реципиент может занести в свою или в заданную (произвольную) базу знаний;

г) Краткое содержание анализируемого текста и т. п.

Результатом работы системы АПТ могут быть разные интеллектуальные продукты (в том числе и перевод на другой язык), свидетельствующие о том, что исходный текст в какой-то мере был понят.

Основной метод КЛ/ВЛ — построение действующих моделей, а затем и компьютерных систем понимания текста. Самая простая схема системы автоматического понимания текста выглядит как преобразование входного текста (Т1), имеющего своего автора, в выходной текст (Т2), понятный адресату:

Автор → Т1 → Компьютер → Т2 → Адресат

ПЛ/ВЛ/КЛ нуждается в собственной теории, объясняющей эти процессы.

§ 2. Что значит «автоматическое понимание текста»

Прежде чем рассматривать основные этапы (или уровни) некоторой действующей модели понимания естественного текста, попробуем определить, что значит *ЭВМ поняла текст*. Мы определим его через результат, который может или стремится построить компьютер.

Результат должен быть другим объектом, отличным от входного текста, т. е. Т2 отличается от Т1 (см. выше). Так, если мы ввели какой-то текст с клавиатуры в ЭВМ, а затем распечатали его на принтере, подключенном к этой ЭВМ, мы не считаем, что текст понят. Но если мы ввели текст на одном языке (скажем, английском), а ЭВМ после работы системы выдала текст на другом языке (например, русском), то можно уже говорить о понимании. Правда, тут важна степень понимания: если машина выдала такой нечленораздельный текст по-русски, который называют абракадаброй, мы скажем, что она «ничего не поняла». Качество результата должен оценивать человек, а оценка всех промежуточных структур — внутреннее дело самой Системы.

Но пока отвлечемся от качества конечного продукта, будем считать, что мы получаем приемлемые результаты, содержательно понятные адресату.

О каких еще результатах машинного понимания, кроме автоматического перевода, можно говорить сейчас?

Будем считать, что машина поняла входной целый текст, если в результате она выдала сжатое изложение (реферат) этого текста. Машина поняла текст, если она может отвечать на вопросы к этому тексту. Машина поняла текст (например, описывающий какую-то картинку или схему), если она может по описанию нарисовать эту картинку либо схему. Машина поняла текст, если она может описанные в тексте сведения (например, о занятости населения нашего города) изобразить в другой форме, например в форме таблицы. Машина поняла текст, если она может сравнить содержание двух разных текстов и сообщить, что в них общего и чем они различаются. Машина поняла текст, если в результате анализа одного, а тем более разных текстов она смогла извлечь такие знания, которые можно поместить в некую копилку человеческих знаний (базу знаний).

Суммируем названные виды понимания:

На входе (Т1)		**На выходе (Т2)**
Естественный текст	⟶	1) текст перевода 2) реферат 3) ответы на вопросы 4) рисунок или таблица
Рисунок или таблица	⟶	текст
Несколько текстов	⟶	результат сравнения: общее/разное фрагмент базы знаний

Чтобы реализовать любой из этих видов понимания, нужно задать много исходных сведений: в виде лингвистических знаний, а также знаний предметных, принадлежащих той предметной области, к которой относится текст. Прежде всего лингвист должен подготовить исходные данные (словари, грамматики, алгоритмы) и сформулировать их так, чтобы их можно было перевести на язык программ, выполнение которых приведет к желаемому результату. Сообщить эти знания компьютеру, который понимает только язык машинных кодов, может программист, который выполняет функцию переводчика, создавая тексты программ. ЭВМ «поняла» текст программы, если она начинает выполнять заданные в программе действия, приводящие в итоге к хорошему или приемлемому Т2.

В одном цикле перехода от входного текста к любому из перечисленных нами возможных результатов содержатся десятки таких преобразований.

Нас интересует, в какой мере за всеми этими переходами стоит единый механизм. Ведь чтобы придти к пониманию в конце цепочки, необходимо, чтобы на всех переходах (от уровня к уровню) также достигалось «понимание». Эту задачу отнесем к компетенции **теории АПТ**.

§ 3. Основные задачи и классы систем АПТ

Вычислительная лингвистика еще не устоялась как научная дисциплина; ее методы, приемы, понятия оттачиваются в многочисленных опытах создания разных систем АПТ.

С одной стороны, машинные ресурсы сейчас так велики, что позволяют хранить в компактном виде все то, что человечество накопило в текстовом виде. С другой стороны, далеко не все тексты заслуживают того, чтобы их хранить в оригинале, и вычислительные способности ЭВМ, а также интеллектуальный потенциал человека открывают возможность преобразовывать тексты совсем в другой вид, не эквивалентный исходному, но зато или более краткий (аннотации, рефераты), или более удобный для дальнейшего использования в формальных системах (например, в форме баз данных, баз знаний).

Конечно, в каждой культуре существует огромный класс художественных произведений, которые являются национальным богатством и которые всегда хранятся полностью, образуя ***полнотекстовые базы данных***. Так, в Институте русского языка Российской академии наук создан и наполняется Машинный фонд русского языка, где основную часть составляет Фонд художественных произведений [см.: Машинный фонд..., 1986]. Чтобы получить сведения о том, каков словарный запас каждого автора, какие у него излюбленные слова и выражения, какие слова он ввел в обиход и в каких контекстах они встречаются, какие употребляемые им слова вышли из обихода, и другие подобные сведения, не нужно просматривать все тексты подряд, достаточно запросить ***информационно-поисковую систему***. Работу такой системы в значительной мере обеспечивает ВЛ, выполняя довольно простые функции, в основном морфологического уровня (например, сводит все словоформы одного слова к его основному, словарному виду для последующего подсчета, сравнения и др.). Но если бы мы захотели сравнить два произведения по содержанию, потребовалась бы гораздо более сложная система, таких систем АПТ пока не существует.

Есть и другие виды текстовых источников, которые требуют хранения полнотекстовых массивов. Это, например, все распорядительные документы (указы президента и постановления и распоряжения правительства) какого-либо государства, а также вся юридическая документация (законопроекты, законы, постанов-

ления и т. п.). Но именно необходимость постоянно обращаться к таким источникам ставит перед ВЛ более серьезные задачи, чем простое хранение.

Задача ведения и поиска в таких массивах лежит на ***информационной системе***: она должна уметь быстро найти все документы, в которых, например, рассматривается тема снятия с должности лиц высокого ранга или тема налогообложения на сверхприбыли и т. п. Такие задачи, как тематический анализ, решаются ***системами автоматического индексирования и рубрицирования***. Первые создают самые простые информационные структуры, называемые поисковым образом документа (ПОД); вторые относят все тексты массива к рубрикам, заданным как значимые для данной предметной области или для данного типа текстов.

В таком массиве очень важно также быстро находить все те документы, на которые явно или неявно ссылается анализируемый текст. Это означает, что массив должен быть снабжен ***гипертекстовой системой***, хранящей связи между текстами и осуществляющей соответствующий поиск. В имеющихся системах гипертекстовые связи проставляются в основном человеком, но это очень трудоемкая задача, к тому же такой подход страдает субъективностью и разностильностью, поэтому развитые системы АПТ ищут способы автоматического построения гипертекстовых связей. В этой актуальной задаче основная нагрузка ляжет на лингвистический аппарат.

Что касается обработки научно-технической литературы и документации, то здесь возникает много вопросов, относящихся к компетенции систем АПТ. Хранить в машинах все, что создано человеком до сегодняшнего дня, не только очень громоздко, но и не нужно: ведь именно технические сочинения очень быстро устаревают, о них достаточно оставить внешнюю информацию: например, такой-то автор писал на такую-то тему или сделал такое-то открытие. Еще больше это относится к потокам сообщений общественно-политического характера: их нужно сортировать на разные массивы по общим темам или источникам (регионам), из которых они получены, нужно сжимать содержательную информацию, формализовать записи и помещать в базы данных и знаний, откуда система будет извлекать и выдавать ответы по запросам.

Таким образом, в задачу автоматической обработки текстов входит и задача автоматического **сжатия** текстовой информации. Ее выполняют ***системы автоматического аннотирования и реферирования***. Этот класс информационных задач (не квазиреферирование, а смысловое сжатие текстов) значительно труднее названных выше, он требует глубокого лингвистического анализа документа, который должен выявить в конечном счете наиболее информативные, наиболее важные части содержания текста. А это уже основная область интересов ВЛ. До настоящего решения такой задачи (т. е. до работы на произвольном корпусе текстов) ВЛ

еще не доросла, эта задача относится к компетенции сложных интеллектуальных систем АПТ.

Системы ***искусственного интеллекта*** (ИИ), работающие с текстовым материалом, опираются на такие компоненты, как базы данных и базы знаний. Они могут задаваться заранее, искусственно вводиться человеком и затем использоваться в автоматическом режиме анализа текста, построения выводов, рекомендаций и т. д. Так они задаются в различных экспертных системах (например, в медицинских диагностических системах), которые могут опираться на сильную формальную логику, но, как правило, не используют лингвистический анализ. Другая разновидность систем ИИ работает с текстом как источником определенных предметных знаний, которые должны быть извлечены в ходе автоматического лингвистического анализа и собраны в структуры баз знаний. Системы такого типа обычно имеют дело с ограниченным корпусом текстов. Эта задача безусловно относится к компетенции интеллектуальных систем АПТ.

С автоматически построенной структурой текстовых знаний связана задача автоматической генерации выходного текста. ***Системы генерации текста*** — наиболее распространенный сейчас тип систем из класса АПТ. По этой парадигме строятся в настоящее время многие системы ***машинного перевода*** — это системы класса «МП на основе знаний». Лингвистические модели таких систем используют весь арсенал собственно лингвистических средств, а также заставляют разработчиков в срочном порядке решать задачи, которые встали перед лингвистикой впервые (в основном это стыковка с предметными областями).

Итак, вот перечень классов систем, содержательным центром которых является автоматическая обработка текста.

1. Хранение текстов. Полнотекстовые базы данных и интеллектуальный поиск.
2. Системы автоматического индексирования и рубрицирования.
3. Системы автоматического аннотирования и реферирования.
4. Информационно-поисковые системы (ИПС).
5. Системы машинного перевода.
6. Системы класса «Искусственный интеллект» (Текст → База знаний).
7. Системы генерации текста (База знаний → Текст).

§ 4. Типы текстовых структур в системах АПТ

Общение на естественном языке заложено в том или ином виде во все современные системы класса «искусственный интеллект» — экспертные, системы общения с банками данных, системы ма-

шинного перевода и др. [см.: Виноград, 1976; Попов, 1982; Мальковский, 1985 и др.]. В справочнике «Искусственный интеллект» они названы «естественно-языковыми системами» [см.: Искусственный интеллект, 1990]. Многие из них работают не только с отдельными предложениями, но и с их объединением, получившим новое качество «связный естественный текст». Процесс анализа в этих системах должен заканчиваться построением семантической структуры, в которой по идее фиксируется «смысл текста». Коротко охарактеризуем некоторые структуры и вклад каждой из них в теорию понимания текста:

А. Лингвистические структуры предложений текста (локальное понимание).

Б. Семантические сети целого текста (глобальное размытое понимание).

В. Информационные структуры целого текста (глобальное обобщенное понимание).

Г. Структуры баз данных и знаний (выборочное специальное понимание).

Д. Структуры систем машинного перевода (параллельное многоязыковое понимание).

Структуры типа А (лингвистические структуры предложений текста) фиксируют результат «буквального» локального (т.е. ограниченного пределами предложения) понимания. Лингвистические процессоры, опирающиеся на сложные и богатые словари, стремятся к полноте интерпретации каждого отдельного предложения исходного текста, к сохранению всей сколь угодно подробной информации о единицах и связях в пределах предложения. В основе лингвистических моделей лежит синтаксическое (или синтактико-семантическое) представление предложения. Будем считать ***классическими лингвистическими*** те семантические структуры, в основе которых лежит лингвистическая модель, или теория, «Смысл ⇔ Текст» [см.: Мельчук, 1999] (далее МСТ или ТСТ).

Главное достоинство лингвистических структур — детальность анализа, отражаемая в форме дерева: ***синтаксического*** и в идеале ***семантического*** представления предложения (далее СинП и СемП соответственно). При наличии словарных статей для всех слов предложения и при условии, что заданная на входе цепочка слов является правильным предложением входного языка, модель и основанная на ней система автоматического перевода строят правильную синтаксическую структуру, сначала поверхностную, затем глубинную (ГСС). Если все узлы ГСС заменить их толкованиями из словаря (например, узел *ПОИТЬ* — на поддерево «***X*** непосредственно каузирует ***Y*** пить ***Z***»), оставив все связи из ГСС и только слегка изменив нотацию (актантные 1, 2, 3, 4; 5 — ATTR, 6 — COORD), мы получим семантическую структуру. Благодаря симметричности процессов анализа и синтеза в МСТ можно идти

от Смысла к Тексту (как в книге И. А. Мельчука), т. е. от заданного СемП к его разверткам в ГСС и далее (см. приложение 1).

В варианте МСТ, реализованном в системе ЭТАП-2 [см.: Апресян, Богуславский, Иомдин и др., 1989], авторы отказались от двухуровневого синтаксиса; единая синтаксическая структура (авторское сокращение — СинтС) помещает в узлах слова исходной фразы и сохраняет подробные связи поверхностной структуры (см. приложение 2). Эти связи, даже если они и имеют грамматический характер, достаточно дифференцированные, и их можно перевести в семантический план, тем более что все они бинарны, а соединяемые ими слова снабжены семантическими характеристиками в комбинаторном словаре.

В названных подходах получающуюся структуру — СемП, ГСС или СинтС — правильнее было бы квалифицировать как **синтактико-семантическое представление** (СинСемП), поскольку ее **основой остается синтаксическое дерево** предложения, но в первом случае имеющее «семантичные» узлы, во втором — «семантичные» связи[1].

Неоспоримым достоинством классических лингвистических моделей является возможность сопоставлять любому предложению обрабатываемого текста его ***формальный структурный образ***. При этом структуры сохраняют всю информацию исходного объекта для дальнейшей автоматической обработки. Благодаря заложенному в модель и систему формальному аппарату описания лингвистических сущностей обеспечен воспроизводимый результат. Излишне говорить, насколько мощный импульс это дает исследованиям в теоретической лингвистике.

Но сейчас нас интересует теория прикладного понимания. С этой точки зрения лингвистические единицы оказываются достаточно жесткими (это древесные структуры), в них можно моделировать лишь понимание в пределах предложения, понимание без обобщений, в котором нельзя опустить ни одно звено. Эти структуры не допускают и характерного для естественного процесса понимания выборочного подхода, «выхватывания» лишь интересующей человека части содержания. «Чисто лингвистическое» понимание является необходимым, но только первым шагом понимания целого текста. Реализация такой идеальной лингвистической модели крайне трудоемка, даже если система предполагает очень ограниченное понимание естественного текста.

Другое узкое место лингвистических моделей — слабая корреляция с единицами представления знаний. Не исключено, что для

---

[1] В приложении 1 воспроизведены примеры СемП одной русской фразы и двух из соответствующих ей ГСС [см.: Мельчук, 1999, *303—306*]. В приложении 2 приводятся СинтС одной английской фразы и вариант соответствующей ей нормализованной русской СинтС [см.: Апресян, Богуславский, Иомдин и др., 1989, *152—153*]. Все пояснения читатель найдет в оригиналах.

общения с каждой конкретной базой данных или для записи в БД информации из текста потребуется создавать отдельную систему перевода.

А какой реальный вид имеет СемП целого текста в чисто лингвистических системах? **Пока реально достижимое СемП целого — это последовательность СинСемП всех подряд предложений текста.**

Если же зафиксировать в СемП сведения о теме-реме (в приложении 1 соответственно ⊥ и ⊻), об эмфатических акцентах и т. п. или даже установить хотя бы только референтные связи между структурами соседних предложений, мы выходим в **структуру типа Б (семантическую сеть целого текста)**. Переведя СинСемП всех предложений текста на язык более элементарных единиц (как предлагается в МСТ, и именно результат такого перевода, его СемП, объявляется «смыслом» текста), мы получим сеть, глобальную «размытую» структуру.

В работе Н. Н. Перцовой утверждается, что модель понимания текста должна включать наряду с поверхностно-семантическим компонентом и глубинно-семантический компонент со своим представлением (ГСемП), объединяющим информацию собственно языковую и энциклопедическую (= общность сведений о действительности, которые имеются у отправителя и получателя текста) [см.: Перцова, 1980]. Это не могут не учитывать системы типа «вопрос-ответ».

В приложении 3 воспроизводится пример несколько упрощенной структуры ГСемП для текста, состоящего из пяти фраз. Способ изображения узлов и связей приближается к принятому в МСТ (узлы более содержательны и эксплицитны, чем связи). Работа с таким объектом оказывается пока практически невыполнимой даже в рамках одного предложения. Это естественно: в глобальном семантическом пространстве целого текста действуют другие законы, чем в пределах предложения.

Рассмотрим другой путь построения «смысла» текста — путь «сверху вниз». Он реализован и реализуется сейчас в разных информационных системах.

Структуры типа В (информационные структуры целого текста) фиксируют результат глобального понимания текста и потока текстов в единицах терминологии выбранной предметной области. Термины сосредоточены в источниках, задаваемых отдельно от текста: классификаторах, тезаурусах, рубрикаторах и др. (см. гл. 8—9). Работающие с этими структурами системы автоматического индексирования и на их основе информационно-поисковые системы имеют дело с реальными текстовыми информационными массивами. Исходный текстовый материал подвергается сжатию: лексический материал текста, не совпавший с единицами тезауруса, просто отбрасывается. Результирующие структуры выражаются в единицах знаний, релевантных для пользователя: дес-

крипторы, термины информационно-поискового тезауруса (ИПТ). Информационный тип моделей — достаточно гибкий:

а) с точки зрения входного анализируемого материала (принимаются естественные тексты без каких-либо структурных ограничений и с довольно большим тематическим разбросом);

б) с точки зрения выходных структур, называемых поисковым образом документа: обычно ПОД представляет собой свободную структуру, элементами которой являются слова ЕЯ, в основном термины, т. е. такие единицы, которые приняты и в естественном понимании. Они не такие мелкие, как в лингвистических моделях, и не такие крупные и подчас отражающие специфическую действительность, как в системах ИИ, они «средние» и обычно имеют переводные эквиваленты в других ЕЯ;

в) с точки зрения средств анализа они не привязаны к конкретной тематике, а достаточно универсальны (так, в некоторых ИПС реализован бессловарный статистический анализ лексического материала, строящий ПОД для текстов любой тематики).

Очень важным параметром ИПС являются реальные, не игрушечные масштабы, наличие компонента «реальный пользователь», с которым у системы может быть организована обратная связь, а это является началом обращения информации. Наличие пользователя (хоть и вне системы) создает реальные критерии оценки качества работы ИПС, что влияет на изменение параметров, учитываемых в работе ИПС. Эти параметры ИПС необходимо учитывать при создании моделей понимания произвольного текста. Узким местом таких систем является небольшой смысловой потенциал (см. об этом в гл. 9).

Структуры типа Г (структуры баз данных и знаний) еще ближе к задачам пользователей и составляют часть реквизита их производственной деятельности. Их можно назвать специализированными, экстралингвистическими структурами, они отображают часть действительности, являются квазиденотатом, который можно привлекать при анализе естественного текста как дополнительное знание. Структуры баз данных — это формальные, жесткие, фиксированные структуры (например, таблицы с описанием кадрового состава учреждения, таблицы занятости населения и др.), поэтому над ними возможны формальные, математически обоснованные операции. Над структурами типа БД можно надстроить лингвистическую систему, генерирующую текст на ЕЯ (см. гл. 10).

Среди них есть и полужесткие структуры динамического типа. Это сценарии, схемы, ситуации или их части — «фреймы». Они образуют базы знаний (БЗн) системы. Такие структуры получили широкое распространение в системах класса ИИ, они отображают сюжет целого текста, который, как известно, безразличен к членению на предложения. При использовании их в качестве инструмента анализа задаваемая жестко схема сюжета дополняется

лингвистическими спецификациями, позволяющими извлекать из текста и вносить в БЗн недостающую информацию.

Большинство подобных структур имеют целью узнавание определенного сюжета в тексте [см.: Семантический компонент..., 1982; Шенк, 1980], как, например, в многочисленных системах Р. Шенка. Путь внешнего понимания, опирающегося не на лингвистическую структуру, а на совпадение лексического состава вопроса и текста при заданной тематике диалога, реализовали многие отечественные коллективы [см.: Нариньяни, 1995; и др.].

Достоинством моделей четвертого типа является возможность сравнения содержания анализируемого текста с разными источниками информации, в частности с другими текстами на ту же тему. Задаваемая тема является как бы ***квазиденотатом***, и такой подход называют иногда ***денотативным*** [см.: Новиков, 1983; Файн, 1987]. Семантические структуры, которые стремятся построить подобные модели, называют (и мы будем в дальнейшем называть) ***концептуальными структурами***.

Главным понятием денотативных (концептуальных) моделей является ограниченный естественный язык (ОЕЯ) [см.: Попов, 1982]. Как ни странно, имеются в виду не столько ограничения на язык, сколько ограничения на мир. Так, системы Шенка узнают «свой» сюжет в текстах любой синтаксической сложности, лишь бы текст был полузеркальным отображением тех сюжетов, которые заложены в базы знаний, их более или менее прямым лексическим соответствием. Переход к новой предметной области требует почти полной смены лингвистического обеспечения подобной системы АПТ. Денотативный анализ не будет иметь успеха и в том случае, когда текст, относящийся к нужной предметной области, формирует «денотат», которого еще нет. Экстралингвистические модели могут хорошо работать в отдельных конкретных задачах, ограниченных набором однотипных текстов, они иллюстрируют зависимость понимания текста от предварительных знаний о предмете, но они плохо или никак не моделируют понимание несюжетных текстов, к которым относятся, в частности, научно-технические тексты. Эти последние имеют тонкие различия с точки зрения аппарата систем АПТ: у них одна лексическая основа, одна база знаний, к ним применимы одни и те же логические выводы, но они передают разные мысли и несут подчас противоположную информацию для читателя. Никакие упрощения поверхностно-языковых средств в виде самого простого ОЕЯ не сделают задачу их понимания доступнее.

Структуры типа Д (структуры систем машинного перевода) используются в системах и моделях автоматического перевода (АП), реализующих самую полную цепочку работы с текстом: от входного текста до выходного, принадлежащего другому ЕЯ. Они фиксируют лексические и синтаксические соответствия (и расхожде-

ния) между единицами и структурами двух языков правилами трансформационного типа Т1 → Т2 (см. схему 1). Еще большее расхождение между входными и выходными единицами можно наблюдать в структурах информационно-переводческих систем, если они осуществляют перевод со сжатием, опираясь на базы знаний: единицы Т1 и Т2 принадлежат разным естественным языкам (ЕЯ-разноязычие), отражают разные объемы содержания (информационное разноязычие), могут различаться фокусами внимания (ценностные, прагматические различия). Такого рода системы АП, основанные на знаниях (англ. *knowledge based machine translation systems*) и совмещающие в себе полноту лингвистического анализа текста с механизмами работы со структурами знаний, являются наиболее перспективными системами автоматического понимания текста [см.: Nirenburg, Carbonell, Tomita et al., 1992]. Но в них тоже не найден пока оптимальный вариант совмещения этих разнородных источников знаний. Не предложен также оптимальный или приемлемый вариант языка-посредника (ЯП), который мог бы служить промежуточным звеном при переводе с любого ЕЯ на любой другой. Предлагаемые разными системами варианты ЯП мало отличаются от лингвистических синтаксических структур.

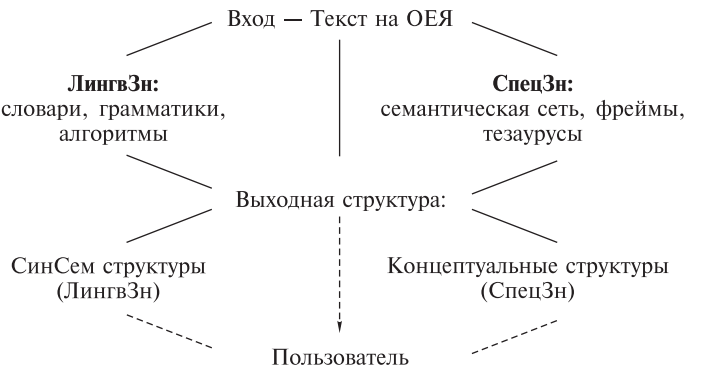
§ 5. Состав компонентов стандартных систем АПТ

Итак, в стандартных системах понимания текста класса ИИ (правильнее было бы говорить об экспериментальных моделях таких систем) на вход поступает текст на ограниченном естественном языке. В качестве средств его анализа (понимания) привлекаются, с одной стороны, лингвистические знания (ЛингвЗн) в виде словарей и грамматик, с другой стороны, специальные знания в выбранной предметной области (СпецЗн) в виде структур ситуаций той предметной области, к которой принадлежит и текст на ОЕЯ. Предметная область, как правило, ограничивается одним сюжетом (например, посещение зубного врача, посещение ресторана, крушение поезда, некоторая политическая ситуация и т. п.). В ответ на конкретный вопрос к тексту или на постоянный информационный запрос (заданный в виде той же схемы предметной области) строится выходная структура конкретного текста, которая и фиксирует результат «понимания» данного текста (схема 1).

Характер системы, т. е. используемые инструменты анализа и тип выходных структур, часто определяется тем, кто был разработчиком — лингвист или специалист в выбранной предметной области. Основой лингвистического подхода является какая-либо синтаксическая модель, а схема предметной области использует-

Схема 1

Упрощенная схема компонентов стандартной системы АПТ



ся стихийно, набегами, в виде фильтров, уменьшающих неоднозначность синтаксического анализа. Выходная структура в терминах ЛингвЗн — это синтактико-семантические представления, которые могут быть основой для перевода (буквального, пофразного) на другой язык; достижим также квазиреферат исходного текста, состоящий из предложений или частей предложений, содержащих нужные слова схемы предметной области.

Для специалиста в предметной области основным средством анализа является семантическая сеть, а морфолого-синтаксические признаки и структуры привлекаются в несистематичном виде: как фильтры, как правила локального контекстного разбора, задачей которого является заполнение оставленных в схеме предметной области пустых мест, т.е. слотов подходящих фреймов. Выходная структура в терминах СпецЗн — это фрагмент семантической сети с заполненными слотами, которую называют уже концептуальной структурой и развертывание которой в текст может достигаться иными средствами, чем использованные в исходном тексте. При этом моделируется не столько понимание, сколько узнавание заданной схемы сюжета, а понимание сводится к поиску в тексте значений некоторых переменных (*Поезд шел из Вены в Париж*, *в результате крушения погибло 10 человек*, *ранено 120* и т.д.). Понимание произвольного текста такие системы не моделируют.

Элементы СпецЗн образуют большую гамму переходов — от атомов смысла (есть словесные атомы, текстовые атомы, у Р. Шенка — спецатомы для каждой узкой проблематики) до самых крупных концептуальных единиц (гиперфреймы, планы, сценарии). Самой популярной единицей словаря в системах АПТ оказался фрейм, так как это средняя единица и по величине, и по глубине, и по разнородности своих дифференциальных признаков.

Стандартным решением проблемы понимания в таких системах АПТ является «ранняя специализация» словарей и грамматик. Это значит, что слова с самого начала вводятся в систему только в тех значениях, которые ожидается встретить в данной узкой области, а грамматики настроены на узнавание лишь тех ситуаций, которые описаны в СпецЗн. Такое решение может привести к тому, что при выходе за пределы уже обработанных текстов система перестает узнавать и свое, и чужое, т.е. и специальное, и общезначимое. Вывод грустный: чтобы угнаться за реальными текстами, нужны все новые и новые словари и грамматики, так что средств анализа окажется больше, чем текстов, которые обрабатываются этими средствами.

§ 6. Модель «мягкого понимания» текста

Охарактеризованные нами несколько типов структур в системах АПТ и способы их построения вовсе не являются конкурирующими или взаимоисключающими. Все они сочетаются в сложном процессе восприятия и понимания текста человеком:

а) пофразное, буквальное, поэлементное чтение и понимание ЕТ;

б) столь же подробное понимание, но в масштабе текста и с отождествлением сущностей, упоминаемых в разных фразах;

в) беглое чтение, выбирающее из ЕТ нужные терминологичные элементы;

г) чтение «крупным взглядом», глазами специалиста, выхватывающего из ЕТ лишь те сюжеты, которые его интересуют;

д) понимание иноязычного текста в единицах своего языка.

Все эти аспекты должны присутствовать в модели понимания произвольного текста произвольным читателем. В этом необходимом наборе приемов понимания отсутствует одно очень важное звено — внутреннее согласование единиц, составляющих сам ЕТ, построение того, что в лингвистике текста называют сверхфразовыми единствами (СФЕ) и межфразовыми связями, и построение на их основе содержательного фокуса текста. Нет той структуры, которую сам ЕТ формирует как свой стержень. Именно он должен «защитить» ЕТ от попыток понять в нем как главное то, что упомянуто между прочим или как пример. Так, нельзя из текста арифметической задачи на вычисление объема воды, вытекающей из бассейна ***А*** в бассейн ***Б***, извлекать информацию об устройстве бассейна. Необходимо вместе с перечисленными приемами понимания ЕТ ввести еще один компонент, вычисляющий фокус ЕТ на основе не внешних, а внутренних свойств единиц текста. Эта функция **информационного синтеза** результатов локального понимания относится скорее к глобальному этапу (см. гл. 7).

Модель «мягкого понимания» текста должна сочетать все эти аспекты, совместить в одной схеме свойства названных структур и механизмов, порождающих их.

Для системы «мягкого автоматического понимания» текста критичной является также проблема «нового», т.е. проблема обработки тех явлений заданного текста, которые описаны неполно или совсем не содержатся в словарях системы. Требуется и адекватная реакция системы на текущий текст, если в нем задана новая информация. Автоматическое накопление этой новой информации для корректировки имеющейся понятийной структуры проблемной области также входит в задачи «мягкой» системы АПТ.

Итак, в модели «мягкого понимания» текста необходимо сочетание двух подходов — ***информационного*** и ***лингвистического***.

§ 7. Синтез информационного и лингвистического подходов

Как результат обобщения, или как теоретическое обоснование разных типов прикладных систем, разработана абстрактная модель, названная **информационно-лингвистической моделью**, в рамках которой можно проследить по шагам все звенья автоматического понимания текста. Коротко, ее смысл сводится к тому, что из одного и того же естественного текста система может извлекать разную информацию для разных пользователей, в зависимости от их интересов, объема знаний, т.е. от того, какие модули компьютерного знания подключены к процессу понимания ЕТ. Такой подход (возможность извлечения разных «смыслов» из единого множества текстов, своего рода лингвистическая относительность) обосновывается и практической необходимостью (действительно, разным пользователям нужна разная информация из текстов), и соображениями конструктивного характера: отдельные части создаются как самостоятельно работающие подсистемы, которые включаются и выключаются в разных режимах работы системы.

Модель мягкого понимания текста, к которому стремятся лингвисты, состоит в способности порождать различные осмысленные интерпретации исходного объекта в зависимости от разных условий и составляющих процесса его восприятия. Модель должна сочетать в себе структуры последовательного, буквального, поэлементного понимания, с одной стороны (узколингвистический подход), и чтение «крупным взглядом», глазами специалиста, с другой стороны (информационный подход, экстралингвистическое понимание).

Концептуальные структуры текста желательно создавать на основе хороших лингвистических представлений текста. Это означает не только учет всех свойств текста от графематических осо-

бенностей и т.д. вплоть до свойства связности текста (это знает лингвист), но и привлечение описаний объектов предметной области в виде тезаурусов или других способов задания специальных знаний (концептуальные структуры, которые умеет строить специалист в данной ПО). Кроме того, концептуальные структуры необходимо соотносить с возможными запросами пользователя (это сфера информатики) и с языком адресата информации (система перевода). Важно также знание того, каковы внутренние установки и цели автора текста (а это прагматика как сравнительно новая область теоретической лингвистики).

Пока никому не удалось реализовать такую модель, но тем более важно рассмотреть с теоретической точки зрения, из каких блоков складывается механизм автоматического понимания.

Главная идея ИЛМ — нащупать новый, плюралистический подход к пониманию ЕТ и примирить методы собственно лингвистического анализа (подробный разбор предложений текста по уровням) и более грубый информационный анализ. Если первый стремится к сохранению и максимальной дифференциации выраженного в тексте содержания, то второй дает взгляд на текст и даже на корпус текстов как на целое, содержание которого может быть представлено с разной степенью подробности или обобщения.

В термин «понимание» вкладывается примерно тот же смысл, что и в термин «информационный анализ»: имеется в виду последовательность операций, приводящих к извлечению из произвольного текста произвольным читателем релевантной информации.

Информационный анализ, понимаемый как построение «текста Информации», неизбежно сопряжен *с **потерей части содержания***. Чтобы эти потери не были случайными, должны быть заданы правила построения наиболее информативных единиц. Такие правила дает лингвистика. Смысл строится только при восприятии текста некоей другой, внешней по отношению к тексту, интеллектуальной системой и зависит от ее установки, ее целей, ее языка и т.д. Те составляющие, которые не работают на смысл, можно «потерять», — этот процесс мы называем «информационным сбросом». Основное назначение лингвистических структур в такой системе АПТ состоит в том, чтобы создавать контекст, необходимый и достаточный для вычленения на каждом уровне информативных единиц, которые переходят в структуры следующего уровня.

Сочетание этих двух тенденций — лингвистической (стремление удержать все) и информационной (стремление утратить, сбросить ненужное) — и есть необходимое условие ***смыслообразования***. Оно не только проявляется на завершающей стадии процесса понимания текста, но и присутствует на всех стыках понимания: например, при переходе от синтаксической структуры к семантической, от семантической к концептуальной и т.д.

Добавление компонента «информационный», усложняя модель в целом, позволит упростить собственно лингвистическую часть, причем не эмпирическим путем, а опираясь на закономерности построения лингвистических объектов и на законы формирования информации.

С практической точки зрения принятый подход (возможность информационного сброса при лингвистическом контроле) позволяет продолжать работу при разных отступлениях от идеальных условий: система может принимать на входе синтаксически неправильные предложения, местоименные и содержащие иные виды неполноты, может работать с неполными словарями, с недоукомплектованными базами знаний и т.д., поскольку какой-то степени понимания можно достичь даже при неполном контексте.

§ 8. Процесс понимания как взаимодействие текстов

Чтобы модель понимания ЕТ, не утрачивая свойств «мягкости», могла стать инженерным (реализуемым на ЭВМ) объектом, необходимо ввести некоторое ограничение на характер компонентов модели. Примем, что все компоненты модели присутствуют в ней в виде текстов и/или структур, переводимых в тексты. Это ***вербальная модель***.

Как известно, в процессе общения участвует довольно большое количество составляющих: сам текст, действительность, говорящий (автор текста), адресат (читатель), коммуникативная среда, база знаний, мотивы и цели говорящего и т.д. Это все разнородные сущности; включить их в одну модель возможно, лишь приведя к одному знаменателю.

В поисках такой однородности ограничимся рассмотрением ***вербального мира*** интеллектуальных систем: составляющие процесса общения будут учитываться в модели лишь в той мере, в какой их можно представить в виде текстов или соответствующих им структур. В модель вовлекается не действительность, а ***текст***, описывающий фрагмент действительности; не знания, а текст, описывающий определенный фрагмент знаний; не цель, а текст, формулирующий цель, и т.д. Понятия «информационная потребность» или «побудительный мотив» могут быть восприняты вербальной системой, если их сформулировать в словесном (или переводимом в словесную форму) виде. Только в этом случае они войдут в систему понимания ЕТ как полноценные составляющие.

Входной объект «естественный текст» обладает наиболее полным комплектом признаков вербальности, или «текстовости» (он имеет автора, размер, композицию, главную тему, входит в какой-то массив и т.п.).

Если принять ограничение по признаку «текстовости», база знаний определенной ПО будет представлена в информационной модели множеством записей, переводимых в высказывания как фрагменты ЕТ, которые можно далее объединять в хорошие или не очень хорошие тексты. Любая порция знания, добавляемая в базу знаний, тоже должна удовлетворять признаку текстовости. Невербализуемый, не переводимый в форму естественно-языкового высказывания фрагмент структуры не будет считаться информацией в вербальной модели.

Знания читателя (пользователя системой АПТ) могут присутствовать в модели в виде множества текстов на разные темы, среди них есть «хорошие» тексты (хорошо структурированные знания), есть «плохие» тексты (им соответствуют обрывочные знания), есть связные, допускающие или не допускающие обобщения, и т.д.

Вопросы, которые пользователь адресует системе, суть фрагменты того или иного «текста его личных знаний». Одно из объяснений «трудных» вопросов состоит в том, что текст вопроса принадлежит одновременно двум разным «текстам знаний».

Множество текстов, привлеченных к процессу понимания некоторого ЕТ, назовем ***информационным пространством*** этого ЕТ. Текст, в терминах которого должен быть представлен результат понимания данного ЕТ, назовем ***встречным текстом***; самый очевидный пример встречного текста — это текст вопроса.

Принятие ограничения на «текстовость» позволяет представить разносортные компоненты системы АПТ как однородные сущности, а их взаимодействие — как ***сравнение и взаимодействие*** обозримых и вполне формализуемых объектов: текстов и их частей (высказываний).

Внутренним строением текстов и их сравнением занимается (или должна заниматься) лингвистика. Поэтому к единицам знаний могут применяться категории оценок, разработанных в лингвистике: например, хорошее знание устроено подобно хорошему тексту или любая порция знания должна быть лингвистически обоснованным объектом. Вытекающие из принципа текстовости следствия имеют содержательный характер и должны выполняться на этапе формирования знаний до того, как будут применяться чисто логические требования строгой однозначности, единственности, непротиворечивости и правильности записей в базе знаний.

Свойство вербальности распространяется и на **минимальные единицы** информационной модели: элементарная единица и текстовых структур, и структур знаний имеет вид R(A,B), что переводимо в простейшее высказывание на естественном языке, поскольку все составляющие этой формулы принадлежат единицам ЕЯ. Свойства этого метаязыка ИЛМ будут рассмотрены в главе 6.

Литература

Анохин П.К. Избранные труды. — М., 1978.

Апресян Ю.Д. Идеи и методы современной структурной лингвистики (краткий очерк). — М., 1966.

Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистическое обеспечение системы ЭТАП-2. — М., 1989.

Баранов А.Н. Введение в прикладную лингвистику: Учеб. пособие. — М., 2001.

Бондарко Л.В., Вербицкая Л.А., Мартыненко Г.Я. и др. Прикладное языкознание: Учебник / Отв. ред. А.С. Герд. — СПб., 1996.

Виноград Т. К процессуальному пониманию семантики // Новое в зарубежной лингвистике. — М., 1983. — Вып. 12.

Виноград Т. Программа, понимающая естественный язык. — М., 1976.

Городецкий Б.Ю. Компьютерная лингвистика: моделирование языкового общения // Новое в зарубежной лингвистике. — М., 1989. — Вып. 24.

Демьянков В.З. Основы теории интерпретации и ее приложения в вычислительной лингвистике. — М., 1985.

Искусственный интеллект: Справочник: В 3 кн. — М., 1990.

Кибрик А.Е. Очерки по общим и прикладным вопросам языкознания. — М., 2001.

Лахути Д.Г. и др. Автоматизированные документальные ИПС: система «Скобки». — М., 1985.

Леонтьева Н.Н. О моделировании «мягкого» понимания текста // Теория и практика общественно-научной информации. — М., 1993. — Вып. 8. — С. 80—97.

Мальковский М.Г. Диалог с системой искусственного интеллекта. — М., 1985.

Машинный фонд русского языка: идеи и суждения. — М., 1986.

Мельчук И.А. Опыт теории лингвистических моделей «Смысл ⇔ Текст». — М., 1999.

Моделирование языковой деятельности в интеллектуальных системах / Под ред. А.Е. Кибрика, А.С. Нариньяни. — М., 1987.

Нариньяни А.С. Проблема понимания ЕЯ-запросов к Базам данных решена // Труды Международного семинара ДИАЛОГ-1995. — Казань, 1995.

Новиков А.И. Семантика текста и ее формализация. — М., 1983.

Новое в зарубежной лингвистике. Компьютерная лингвистика / Под ред. Б.Ю. Городецкого. — М., 1988. — Вып. 24.

Падучева Е.В. Высказывание и его соотнесенность с действительностью. — М., 1985.

Перцова Н.Н. К построению глубинно-семантического компонента модели понимания текста // Проблемы вычислительной лингвистики и автоматической обработки текста на естественном языке. — М., 1980. — С. 3—89.

Попов Э.В. Общение с ЭВМ на естественном языке. — М., 1982.

Поспелов Д.А. Логико-лингвистические модели в системах управления. — М., 1981.

Семантический компонент в системах автоматического понимания текстов. Обзорная информация. — М., 1982. — Вып. 6.

Структурная и прикладная лингвистика / Под ред. А.С.Герда. — СПб., 2004. — Вып. 6.

Файн В.С. Распознавание образов и машинное понимание естественного языка. — М., 1987.

Шенк Р. Обработка концептуальной информации. — М., 1980.

Nirenburg S., Carbonell J., Tomita M. et al. Mashine Traslation: A Knowledge-Based Appach. — Pittsburgh, 1992.

ГЛАВА 2

МАШИННЫЙ ПЕРЕВОД КАК СРЕДА СОЗДАНИЯ СИСТЕМ АВТОМАТИЧЕСКОГО ПОНИМАНИЯ ТЕКСТА

Машинный перевод, или ***автоматический перевод*** (АП), — это интенсивно развивающаяся область научных исследований, экспериментальных разработок и уже функционирующих систем машинного перевода (СМП), в которых основная часть процесса перевода с одного естественного языка (ЕЯ1) на другой (ЕЯ2) выполняется компьютером. СМП призваны обеспечить быстрый и систематический доступ к информации, содержащейся в больших потоках текстов на иностранном языке. Промышленные СМП, переводя в основном научно-технические тексты, опираются на большие терминологические банки данных, поддерживая единообразие в переводе терминологической и специальной лексики. Они обычно требуют привлечения человека в качестве пред-, интер- и/или постредактора.

Помимо практической потребности делового мира в системах МП существуют и чисто научные стимулы к их развитию: экспериментальные СМП являются опытным полем для проверки различных аспектов теории компьютерного понимания, речевого общения, методов преобразования информации, а также для создания новых, более эффективных моделей самого машинного перевода. Современные СМП, использующие базы предметных знаний в качестве промежуточного звена, относят к классу систем искусственного интеллекта.

§ 9. Об истории СМП

В 1954 г. в США был осуществлен знаменитый Джорджтаунский эксперимент по машинному переводу с русского языка на английский. Хотя был задействован маленький словарь (250 слов) и тщательно отобраны фразы для перевода, этот первый опыт обеспечил бурное развитие работ по МП в течение десяти ближайших лет. В 1955 г. прошел первый эксперимент по МП в СССР (англо-русский, со словарем 2300 слов из области прикладной математики, в дальнейшем эти разработки вошли в состав систе-

мы АМПАР). Начались работы по МП в Институте прикладной математики, где созданы три экспериментальные системы (с французского на русский ФР-1 и ФР-2 и с английского на русский) под руководством О.С.Кулагиной и И.А.Мельчука. Заслуга этих ученых состояла еще и в том, что они положили начало формированию теории МП.

В 1959 г. открывается Лаборатория машинного перевода в МГПИИЯ им. М.Тореза (сейчас МГЛУ), где на материале разных языков строятся модели систем МП, создаются словари, разрабатываются концепции семантического языка-посредника. Из масштабных работ созданы лингвистическое обеспечение и словарь для системы англо-русского перевода (АРАП), но машинная реализация лингвистических разработок отложена на многие годы ввиду отсутствия техники.

С 1966 г. работы по СМП в США были приостановлены (как следствие выводов специальной комиссии о нерентабельности МП по сравнению с обычным переводом), переориентированы были и работы в СССР. Но к середине 1970-х гг. интерес к проблеме МП возродился, начинается бурное развитие систем во многих западных странах. Всплеск исследований и работ по СМП наблюдается в Японии и Китае. К концу 1980-х гг. в Японии было уже около трех десятков систем МП, в Китае — около 10, из них две — промышленные.

В США работает в промышленном режиме несколько десятков систем, первая и наиболее известная из них — СИСТРАН — эксплуатируется с 1970 г.

Интенсивно развертываются работы по МП в Гренобле (Франция). Создаются системы GETA (CETA), работающие на основе синтаксического анализа [см.: Boitet, Nedobejkine, 1980], затем СМП ARIANE 78 [см.: Müller, 1983] — это МП с русского языка на французский.

В 1976 г. в Монреале (Канада) начинает работать первая полностью автоматическая система МП TAUM-METEO, переводящая тексты метеосводок с английского языка на французский.

И в России активизировался машинный перевод. Группы машинного перевода возникают почти во всех крупных университетах страны. Многообещающие работы ведутся в Ленинграде, Киеве, Тбилиси, Ереване. В 1975 г. во Всесоюзном центре переводов (ВЦП) начата разработка трех систем промышленного масштаба: англо-русской (АМПАР), немецко-русской (НЕРПА) и французско-русской (ФРАП) — они описаны в книге У.Хатчинза [см.: Hutchins, 1986]. Было сдано в эксплуатацию по две версии каждой системы, а системы АМПАР и ФРАП начали выполнять реальные заказы. С наступлением эры персональной техники и новых экономических условий практически все эти работы были свернуты, а коллективы переориентировались на другие задачи.

В 1974 г. начались работы по МП в ИНФОРМЭЛЕКТРО, в дальнейшем коллектив перешел в ИППИ РАН, где под руководством Ю. Д. Апресяна создано семейство систем ЭТАП (версии 1, 2, 3) — систем МП с французского и с английского языков на русский — они успешно развиваются по настоящее время.

В Институте востоковедения РАН также успешно идут работы по японско-русскому автоматическому переводу (ЯРАП) под руководством З. М. Шаляпиной.

На рубеже XX и XXI вв. разработки ленинградских коллективов (при иностранной поддержке) дали начало отечественной промышленной системе ПРОМТ. В ряде российских фирм начаты работы по МП с турецкого языка на русский и английский.

История МП хорошо документирована. Начало этому положено И. А. Мельчуком и Р. Д. Равич в двух томах полной библиографии работ, связанных с МП как научной дисциплиной, с аттестацией содержательных операций во всех имеющихся к этому моменту системах и разработках. Обзор охватывает период 1949—1970 гг.

В справочнике по системам ИИ приведен перечень около 70 основных систем МП, созданных к концу 1980-х гг., с их параметрами: организация и страна разработки, направления переводов, состояние разработки, основная литература по каждой системе [см.: Искусственный интеллект, 1990]. Все перечисленные СМП кратко охарактеризованы по типу лингвистической стратегии и т. п.

Следующие сколько-нибудь полные обзоры на русском языке по МП нам неизвестны. Летопись СМП вел и продолжает вести У. Хатчинз [см.: Hutchins, 1986]. Частные проблемы МП обсуждаются постоянно в материалах периодических конференций — COLING, Computational Linguistics, ACL-Proceedings, Пражский Бюллетень по математической лингвистике, МЕТА (Канада), ДИАЛОГ и др. Более или менее полные описания СМП содержатся в технических отчетах.

§ 10. Периодизация и классификация СМП

Системы машинного перевода можно классифицировать по нескольким основаниям. Одно из них — принятая в системе лингвистическая стратегия. С точки зрения ее развития выделяется четыре периода.

Начальный период «бурного развития» (до середины 1960-х гг.) характеризуется преимущественным развитием прямых систем МП, обеспечивающих результаты, близкие к пословному переводу. Это системы МП ***первого поколения***. В них операция перевода требует минимума преобразований: исходный текст постепенно

превращается в текст на выходном языке путем замены всех его элементов, найденных в словаре, на переводные эквиваленты. Учет локального контекста позволяет собрать некоторые сложные единицы — обороты, поэтому такой перевод называют еще ***пословно-пооборотным***. Наличие неполного синтаксического анализа относит систему уже к ***полуторному поколению***. Эти СМП ***бинарны***, ***одновариантны***, не имеют промежуточных структур.

Второй период (середина 1960—1970-х гг.) отмечен интенсивным развитием синтаксических теорий и разработкой на их основе СМП ***второго поколения***. В них переводные соответствия устанавливаются не прямым способом, а через построенную для каждого предложения синтаксическую или синтактико-семантическую структуру (или несколько вариантов такой структуры). Анализ и синтез в них независимы: анализ, как правило, ***многовариантный***, ведется в категориях входного языка, синтез — в категориях выходного. Связь того и другого этапов обеспечивается третьим компонентом — этапом ***межъязыковых операций*** **(*трансформаций*)**, это собственно перевод, или ***трансфер***.

Третий период (середина 1970—1980-х гг.) можно назвать периодом экстенсивного развития СМП: они выходят в промышленность. Техника морфологического и синтаксического анализа хорошо освоена, но остро ощущается недостаток семантики. Однако ожидаемого выхода к СМП ***третьего поколения***, который бы осуществлял перевод через семантический язык-посредник, универсальный для разных пар естественных языков, не произошло. Такой путь не был обеспечен единой общепризнанной теорией. На этом лингвистические основания классификации СМП прервались.

В качестве компенсации получают развитие ***интерактивные*** СМП, комбинирующие труд человека и ЭВМ. Другое внешнее решение семантических трудностей — ориентация на перевод ограниченных классов текстов, охватывающих узкую предметную область.

Четвертый период (со второй половины 1980-х гг.) характеризуется возрастанием интереса к МП как с практической, так и с теоретической точки зрения. МП — сложная область, на которой отрабатываются новые информационные технологии. Большие надежды возлагаются на мощные ***лексические*** и ***терминологические базы данных*** и базы знаний. В МП привлекаются семантические теории из узких предметных областей или из экспертных и других систем ИИ.

В отдельный класс выделяются системы МП, основанные на знаниях (*knowledge-based MT*, или *KBMT systems*), создается несколько экспериментальных систем МП, использующих ***интерлингву*** **(*язык-посредник*)** и структуры ***представления знаний*** [см.: Nirenburg, 1989; и др.]. Входят в моду ***концептуальные структуры***,

концептуальные сети. Но этот термин допускает разные понимания; часто структура, названная концептуальной, мало отличается от синтаксической.

По количеству привлекаемых языковых пар СМП делятся на ***двуязычные*** (ориентированные только на данную пару языков) и ***многоязычные***. Те и другие могут быть ***бинарными*** (если анализ входного языка ведется в категориях выходного) или ***универсальными*** (если устройство анализа не зависит от выходного языка). Так, система СИСТРАН многоязычная, но не универсальная, так как состоит из совокупности бинарных СМП.

По тематической ориентации различают системы ***монотематические***, настроенные на одну ПО (таких большинство: TAUM-METEO, METAL, TITUS, SPANAM), и ***политематические***. Некоторые СМП имеют ограничения на структуру вводимых текстов — это системы с *ограниченным ЕЯ*. Так, TITRAN переводит только заголовки.

По степени участия человека можно говорить о ***полностью автоматическом*** переводе и ***человекомашинном переводе***. Представителем первого является система TAUM-METEO, переводящая сводки погоды (в двуязычной Канаде) с английского на французский, — это единственная полностью автоматическая СМП. Сейчас на ее основе работает система FoG, генерирующая тексты на этих двух языках из общей базы данных метеосводок (см. гл. 10).

С точки зрения степени разработанности СМП образуют три класса: ***промышленные***, ***развивающиеся*** и ***экспериментальные***.

Особый класс образуют СМП, основанные на образцах (*example-based*), или прецедентах. Несмотря на заманчивость и кажущуюся легкость реализации такого подхода (текст-текст, шаблон-шаблон), они остаются экспериментом [см.: Михеев, 2004], хотя во многих системах такой компонент так или иначе присутствует.

§ 11. Лингвистическое обеспечение СМП

Процесс МП представляет собой последовательность преобразований, применяемых к входному тексту и его структурам и превращающих его в текст на выходном языке, который должен максимально воссоздавать смысл и, как правило, структуру исходного текста, но уже средствами выходного языка. В классических СМП, осуществляющих непрямой перевод по отдельным предложениям (*пофразный перевод*), каждое предложение проходит последовательность преобразований, состоящую из трех частей, или этапов:

АНАЛИЗ → **ТРАНСФЕР** (межъязыковые операции) → **СИНТЕЗ**

Цель этапа анализа — построить структурное описание (*промежуточное представление*, *внутреннее представление*, обычно это *СинП — синтаксическое представление*) входного предложения. Задача этапа трансфера (собственно перевода) — преобразовать структуру входного предложения во внутреннюю структуру выходного предложения. К этому этапу относятся и замены лексем входного языка их переводными эквивалентами (лексические межъязыковые преобразования). Цель этапа синтеза — на основе полученной в результате анализа структуры построить правильное предложение выходного языка.

Как правило, этапы анализа и синтеза строятся зеркально (см. обобщенную схему МП в приложении 4). Первыми такую схему стали использовать З. М. Шаляпина и Б. Вокуа еще в 1970-х гг. Эта основа варьировалась и детализировалась в разных системах. Анализ проходит цепочку этапов: **ДоСинАн — СинАн** — иногда **СемАн** (или даже **Концептуальный Анализ —** только для узких ПО).

Но это теоретически. Практически же в работающих системах реализован путь от СинП входного предложения до СинП и далее цепочки слов выходного предложения. А поскольку многие проблемы синтаксиса не могут быть решены без привлечения семантики, то в разных системах МП, не имеющих отдельного семантического или концептуального уровней анализа, состав синтаксического компонента сильно разнится (см. приложения 5—9).

Различия могут объясняться способом задания грамматики (в виде словарных статей, или последовательностью правил непосредственно составляющих, или системой алгоритмов и т. д.). Обычно СМП не ограничиваются одним способом, а используют несколько разных способов в разных комбинациях, включая эмпирические правила, особенно этим грешат промышленные системы, о принципах работы которых почти ничего не сообщается (СИСТРАН, ПРОМТ). В экспериментальных системах различают системы, работающие «под управлением» словаря, правил, системы, основанные на синтаксисе, семантике, знаниях, образцах и т. д. (*lexicon-driven*, *rule-driven*, *syntax based*, *semantics based*, *knowledge based*, *example-based*, *etc.*).

Системы МП второго поколения и выше отличает высокая ***модульность***, что выражается в том, что изменения внутри модуля (будь то алгоритм, словарь, грамматика или любое промежуточное представление) не влияют на вид информации, подаваемой на его вход и выдаваемой на его выходе, так что отдельные части грамматик и словарей можно менять и дополнять, не меняя всей системы.

Конкретные соотношения различных модулей системы (словари — грамматики, грамматики — алгоритмы, алгоритмы — программы, декларативные — процедурные знания и др.), включая

распределение лингвистических данных по уровням, — это то основное, что определяет специфику СМП.

§ 12. Внешняя и внутренняя оценка СМП

В том большом эксперименте, который представляют собой СМП, для каждой **полной системы** должны быть сформулированы критерии оценки — ***внешней*** (оценка результатов работы системы пользователем) и ***внутренней*** (оценка самой системой результатов каждого этапа — для экспериментальных систем).

В исследовании О.Кулагиной рассматриваются основные параметры качества, по которым давалась внешняя оценка результатов во многих СМП [см.: Кулагина, 1979]. Это ***понятность*** и ***правильность***, или ***адекватность***, перевода. При пофразной оценке переводов на массиве 15 тыс. слов, выполненных системой GETA (Франция), вполне понятных фраз оказывалось 50%, не совсем понятных — 28%, совсем непонятных — 22%. Разбиение на классы качества достаточно субъективно. В оценке системы ФР-2 учитывались три категории качества: **понятность**, **адекватность** и **грамматическая правильность** — и проводилось разбиение на категории самих экспертов: лингвисты, математики (т.е. специалисты в той ПО, которой принадлежали тексты) и смешанные группы. Оценка усложняется, если система выдает несколько переводов одной фразы.

Возможна и внутрисистемная оценка самой системой результатов каждого этапа. Эту функцию выполняет часто специальный компонент «Аксептор». Самым очевидным случаем самооценки является принцип «все или ничего», т.е. приятие или неприятие следующим этапом результатов предыдущего. Таким жестким принципом долгое время руководствовались системы второго поколения. Если строится дерево предложения, оно поступает на синтез; если же хотя бы на одном участке неудача, вся фраза отбраковывается. В дальнейшем этот принцип был смягчен использованием эвристик, приписыванием весов, выходом на более простой режим перевода (перевод по синтаксическим группам, пословный и т.п.), а также особой организацией грамматики, когда первым выдается самый лучший вариант (например, в системе ЭТАП).

В сборнике «Машинный перевод и прикладная лингвистика» предлагается проводить внутрисистемную оценку (пока теоретически) по основному параметру каждого уровня понимания [см.: МП-271]. Текст перевода должен оцениваться мерой отступления от идеальных требований:

а) ***непрерывность*** — это отсутствие вариантов перевода одной и той же единицы;

б) ***грамматическая правильность*** — отсутствие нарушений законов построения синтаксической структуры выходной фразы;

в) ***семантическая связность*** — отсутствие нарушений правил семантической грамматики и отсутствие разрывов в семантическом графе целого текста;

г) ***осмысленность*** в данной ПО — возможность непротиворечивых интерпретаций структуры целого текста в единицах данной предметной области.

§ 13. Нерешенные проблемы автоматического понимания и перевода

В настоящее время в России существуют и развиваются серьезные (т.е. с полноценным лингвистическим обеспечением) СМП семейства ЭТАП [см.: Апресян, Богуславский, Иомдин и др., 1989] и система ЯРАП [см.: Шаляпина и др., 2001]. Начато создание нескольких «молодых» СМП (с татарского и турецкого языков, по которым появляются робкие заявления и публикации), авторы которых заново проходят уже пройденные пути и снова преодолевают те трудности, которые детально описало старшее поколение. На помощь им приходят новые информационные технологии, огромные размеры машинной памяти, корпусная лингвистика. Но с помощью только информационных технологий невозможно решить принципиальные узловые проблемы автоматического понимания текста. Главные проблемы здесь следующие:

а) решение ***неоднозначности*** формального синтаксического анализа изолированных предложений текста; считается, что формальная неоднозначность решается выходом к семантической интерпретации; если таковой нет, синтаксический уровень нужно строить так, чтобы первым выдавался самый правильный вариант — он и поступает на перевод;

б) преодоление структурной и смысловой ***неполноты*** локальных участков (предложений) текста; видимо, она может быть восполнена выходом к межфразовому анализу; для перевода это в первую очередь проблема правильного восстановления антецедентов местоимений;

в) организация ***гибкого*** подключения разных предметных областей к процессу понимания и перевода; эту проблему предлагается решать организацией общей онтологии в сочетании с частными базами знаний;

г) необходимость понимания текста как ***целого образования*** (в противоположность псевдопониманию изъятых из него частей). Это необходимо для правильного перевода заголовков, подписей под схемами и другими изолированными частями текста. Решение

этой проблемы необходимо и для реализации систем перевода класса KBMT.

Эти проблемы нужно решать в комплексе, заранее определяя место для каждой из них в сложной архитектуре системы. При отсутствии общего проектирования названные проблемы будут возникать при анализе каждого отдельного предложения и должны будут решаться эмпирически (т.е. требовать немалых затрат и «заплат»).

Первой самой необходимой задачей является поиск способа включения знаний предметной области в лингвистические процессоры.

§ 14. Новая парадигма СМП

Мы считаем, что именно осознание факта несимметричности процесса понимания текста привело разработчиков систем МП к смене традиционной парадигмы (в которой анализ и синтез трактуются как симметричные, хоть и достаточно независимые процессы) на несимметричные типы систем. В МП стали преобладать системы перевода через базы знаний — KBMT (Knowledge-Based Machine Translation), вместо компонента синтеза появились системы типа ***генераторы текстов*** [см.: Nirenburg, 1989; и др.]. В новой парадигме обострилось внимание к специальным знаниям, к прагматическому аспекту знаний и т.д., что помогает уточнить само понятие семантической структуры текста, разделить понятия семантической и концептуальной структур. К компоненту синтеза, или генерации, текста стали предъявлять требования правильно построенного дискурса независимо от того, каков источник информации: семантическое представление ранее проанализированного текста, протокол машинного эксперимента, запись в базе данных или даже нетекстовый объект из арсенала средств Multimedia [см.: Roesner, 1987; McKeown, 1988].

Задачи АПТ в свете этой новой парадигмы можно обобщить так:

1. Анализ исходного ЕТ, который обеспечивает построение лингвистических структур, в том числе разных семантических структур, полных, частичных, сжатых, стремящихся представить содержание текста в форме баз данных, в виде концептуального графа и т.д.

2. Сравнение лингвистических структур текста со специальными или с индивидуальными знаниями, также представленными в форме БД.

3. Генерация текстов на основе информации, заключенной в традиционных реляционных БД, а также в концептуальных текстовых структурах или в индивидуальных базах знаний.

§ 15. Включение предметной области как задача информационно-переводческой системы

За подключение любого компонента специальных знаний к процессу понимания исходных естественных текстов ответственна лингвистика, это она должна обеспечивать общение друг с другом «разноязычных» систем: ведь очень непросто включить в лингвистические структуры экстралингвистические знания, имеющие совсем другую природу. Правда, в вербальную систему допускаются только словесно оформленные знания, но в каждом случае система должна преодолевать информационное разноязычие: в пределах одного языка, в нашем случае русского, система понимания должна уметь преобразовывать друг в друга объекты разной природы — лингвистические и специальные.

Именно поэтому за основу модели «мягкого» (адаптируемого к разным ПО) понимания должна быть взята не просто модель систем автоматического перевода, реализующих самую полную цепочку работы с текстом, а схема информационно-переводческих систем. Отказываясь от задачи полного, точного, со всеми лингвистическими нюансами перевода, эти системы не отказываются от задачи извлечения и перевода информации, заключенной в тексте.

Язык, удовлетворяющий как задачам общения текста с разными базами данных, так и задачам перевода с одного ЕЯ на другой, должен быть ***информационным языком-посредником***. Тогда информационно-переводческие системы смогут осуществлять перевод со сжатием и с использованием баз знаний; при этом входные и выходные единицы (тексты) могут принадлежать разным естественным языкам, отражать разные объемы содержания, могут различаться структурами, фокусами внимания и т. д. Структуры выходного языка оказываются в этом случае воспринимающей системой, т. е. необходимым компонентом информационного анализа. В приложении 9 схематически показано совмещение функций системы МП, привлечения специальных знаний и сжатия текста.

* * *

До сих пор не создан ни один проект МП, дающий удовлетворительное качество перевода. Однако современные системы МП, основанные на детальном, полном анализе текста, содержат в себе большой потенциал, они в принципе ***многофункциональны***. Пути развития систем МП характеризуются большим разнообразием. Одни системы начинаются с собственно системы машинного перевода и постепенно образуют среду для создания любых прило-

жений — не только МП, но и обучающих систем, средств аннотирования (пока морфо-синтаксического) больших корпусов текстов, автоматического редактирования, посредника для общения с базами данных типа SQL, информационного поиска и т.п. Таковы цели, которые ставятся сейчас перед лингвистическим процессором ЭТАП-3 [см.: Apresian, Boguslavsky, Iomdin et al., 2003]. По-видимому, коллектив связывает большие надежды с переходом на сетевой язык межнационального общения UNL. Другие проекты сразу планируются для широкого круга задач. Такова система ПОЛИТЕКСТ, которая получила к настоящему моменту продолжение в виде двух систем: информационной системы РОССИЯ и системы МП с русского языка на английский ДИАЛИНГ. Третьи системы сосредоточены на идее перевода через язык-посредник и базы знаний и развивают методы организации онтологических знаний, общих для разных языков [см.: Nirenburg, Carbonell, Tomita et al., 1992]. В системах МП, работающих в реальном масштабе, главное внимание уделяется средствам автоматизированного создания больших терминологических банков. В системах, создаваемых в последнее время, возродился интерес к использованию параллельных корпусов — источника готовых переводов словосочетаний, трудных для синтаксического анализа. Отметим и неослабевающий научный интерес к поиску более экономной организации комплекса словарей как пока основного источника информации, необходимой для работы систем АПТ. В работах З.М.Шаляпиной и С.А.Крылова развиваются более реалистичные идеи о путях совершенствования систем МП [см.: Шаляпина, 1996; Крылов, 2004]. Что касается второй работы, то она предлагает гибкое сочетание человеческого и машинного методов с постепенным обучением системы МП. М.Нагао видит будущее МП в создании частных систем [см.: Синтаксический компонент..., 1981]. Этот путь мы далее не рассматриваем, так как нас интересуют теория и механизмы создания лишь полностью автоматических систем.

Литература

Автоматический перевод: Сб. статей / Под ред. О.С.Кулагиной и И.А.Мельчука. — М., 1971.

Актуальные вопросы практической реализации систем автоматического перевода // Материалы Первого совместного советско-французского семинара, состоявшегося в Москве в 1977 г.: В 2 ч. — М., 1982.

Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистическое обеспечение системы ЭТАП-2. — М., 1989.

Иомдин Л.Л. Уроки русско-английского (из опыта работы системы машинного перевода) // Труды Международной конференции ДИАЛОГ-2002. — М., 2002. — Т. 2. — С. 234—244.

Искусственный интеллект: Справочник: В 3 кн. — М., 1990.

Крылов С.А. Обучаемость системы АП как основа осуществимости безошибочного автоматического перевода, или О переходе количества в качество. — М., 2004.

Кулагина О.С. Исследования по машинному переводу. — М., 1979.

Леонтьева Н.Н. Система французско-русского автоматического перевода (ФРАП): Лингвистические решения, состав, реализация // МП-271. — М., 1987. — Вып. 271. — С. 6—25.

Леонтьева Н.Н., Никогосов С.Л. Система ФРАП и проблема оценки качества автоматического перевода // Машинный перевод и прикладная лингвистика. — М., 1980. — Вып. 20. — С. 57—78.

Машинный перевод и прикладная лингвистика. — М., 1958—1985.

Мельчук И.А. Опыт теории лингвистических моделей «Смысл ⇔ Текст». — М., 1999.

Мельчук И.А., Равич Р.Д. Автоматический перевод (1949—1963): Критико-библиографический справочник. — М., 1967.

Михеев М.Ю. Перевод на основе базы прецедентных словосочетаний, или переводных фрагментов // Труды международной конференции ДИАЛОГ-2004. — М., 2004.

МП-271 — Машинный перевод и прикладная лингвистика. Проблемы создания системы автоматического перевода: Сб. науч. трудов МГПИИЯ им. М.Тореза. — М., 1987. — Вып. 271.

Проблемы анализа и синтеза целого текста в системах машинного перевода, диалоговых и информационных системах. Обзорная информация / Сост. С.И.Гиндин, Н.Н.Леонтьева. — М., 1978.

Синтаксический компонент в системах машинного перевода. Обзорная информация / Сост. Н.Н.Леонтьева, З.М.Шаляпина и др. — М., 1981. — Вып. 5.

Слокум Дж. Обзор разработок по машинному переводу: история вопроса, современное состояние и перспективы развития // Новое в зарубежной лингвистике. Компьютерная лингвистика. — М., 1989. — Вып. 24. — С. 357—408.

Чернов Г.В. Основы синхронного перевода. — М., 1987.

Шаляпина З.М. Автоматический перевод: эволюция и современные тенденции // ВЯ. — 1996. — № 2. — С. 105—117.

Шаляпина З.М. Макет лингвистического обеспечения системы японско-русского АП ЯРАП. — М., 1980.

Шаляпина З.М. и др. Экспериментальный комплекс ЯРАП для лингвистических исследований в области японско-русского автоматического перевода: первая очередь. — М., 2001.

Apresian J.D., Boguslavsky I.M., Iomdin L.L. et al. ETAP-3 Linguistic Processor: A Full-Fledged NLP Implementation of the Meaning ⇔ Text Theory // First International Conference on Meaning-Text Theory. — Paris, 2003. — P. 279—288.

Bernard Vanquois et la TAO. Vingt-cing ans de Traduction Automatique: Analectes / Ed. Ch. Boitet. — Grenoble, 1989.

Boitet Ch., Nedobejkine N. Russian-French at GETA: Outline of the Method and Detailed Example // Proceedings of the 8-th International Conference on Computational Linguistics. — Tokyo, 1980. — P. 437—446.

Hutchins W.J. Machine Translation: Past, Present, Future. — New York, 1986.

Isabelle P., Bourbeau L. TAUM-AVIATION: Its Technical Features and Some Experimental Results // Computational Linguistics. — Grenoble, 1985. — Vol. 11. — № 1. — P. 18—27.

King M. (éd.), Machine translation today: the state of the art. — Edingurgh, 1987.

Kittredge R. The significance of sublanguage for automatic translation. — Pittsburgh, 1987.

McKeown K. Text Generation. — Cambridge, 1988.

Mel'c'uk I.A., Ravic' R.D. Traduction Automatique (1967—1970) / Ed. A.V. Gladkij. — Montreal, 1978.

Müller A. ARIANE-78, système de traduction assistée par ordinateur. — Paris, 1983. — V. 21. — № 3.

Nirenburg S. Knowledge-Based Machine Translation // Machine Translation. — Pittsburgh, 1989. — № 1.

Nirenburg S., Carbonell J., Tomita M. et al. Machine Translation: A Knowledge-Based Approach. — Pittsburgh, 1992.

Roesner D. The Generation System of the SEMSYN Project: Towards a task — independent Generation for German // 1st European Workshop on Language Generation. — 1987.

Slocum J. A Survey of Machine Translation: Its History, Current Status, and Future Prospects // Computational Linguistics. — Grenoble, 1985. — Vol. 11. — № 1.

Sumita E. and Iida H. Experiments and Prospects in Example-Based Machine Translation // Proceedings of ACL-91. — P. 185—192.

Vauquois B., Boitet Ch. Automated Translation at Grenoble University // Computational Linguistics. — Grenoble, 1985. — Vol. 11. — № 1. — P. 28—36.