ЯЗЫКОЗНАНИЕ

# КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

Ю. Н. Марчук

*Учебное пособие*

ВОСТОК
ЗАПАД

**Ю. Н. Марчук**

# КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

*Учебное пособие*

Рекомендовано Мииистерством образования и науки Российской Федерации в качестве учебника для студентов высших учебных заведений, специализирующихся по направлению и специальности «Филология»

Москва
АСТ ВОСТОК ЗАПАД
2007

УДК 811.93(075.8)
ББК 81.1я73-1
М 30

Подписано в печать 07.06.06. Формат 84x108 1/32.
Усл. печ. л. 16,8. Тираж 1 000 экз Заказ № 6888

**Марчук, Ю.Н.**

М30 Компьютерная лингвистика : учебное пособие / Ю.Н. Марчук. — М.: АСТ: Восток —Запад, 2007. — 317, [3] с.

ISBN 5-17-039480-2 (ООО «Издательство АСТ»)
ISBN 5-478-00383-2 (ООО «Восток — Запад»)

Учебник посвящен лингвистическим основам обработки текстов на естественном языке посредством компьютера. Рассмотрены ввод в компьютер и обработка лингвистической информации, терминология и терминография, моделирование, экспертные системы, машинный перевод.

Предназначен для студентов и аспирантов, всех интересующихся проблемами современной лингвистики и ее приложения.

УДК 811.93(075.8)
ББК 81.1я73-1

**ISBN 5-478-00383-2**

# Оглавление

## Об авторе

МАРЧУК Юрий Николаевич — доктор филологических наук, профессор, академик Международной академии информатизации, заведующий кафедрой теоретической и прикладной лингвистики Московского государственного областного университета, профессор кафедры общего и сравнительно-исторического языкознания филологического факультета Московского государственного университета им. М. В. Ломоносова, профессор кафедры экспериментальной и прикладной лингвистики Московского государственного лингвистического университета.

Автор более 320 печатных работ, среди которых книги, вышедшие в издательстве «Наука»: «Проблемы машинного перевода», «Методы моделирования перевода», учебные пособия «Вычислительная лингвистика», «Автоматический контекстологический словарь для машинного перевода многозначных слов с английского языка на русский», в качестве ответственного редактора руководил выпуском многоязычных словарей лингвистических и географических терминов.

Читал лекции и вел научную работу в Камбодже, Лаосе, Корее, США, Франции, Китае, участвовал в целом ряде международных научных конференций в нашей стране и за рубежом.

Координировал разработку систем машинного перевода в СССР, руководя работой Всесоюзного центра переводов научно-технической литературы и документации Государственного комитета СССР по науке и технике и Академии наук СССР. Непосредственно участвовал и руководил исследованиями и разработками систем машинного перевода АМПАР, СПРИНТ, НЕРПА и др. Награжден медалью ВДНХ СССР за эти разработки. В настоящее время руководит выпуском серии многоязычных терминологических словарей по лингвистике, географии, истории. За многолетнюю и плодотворную педагогическую деятельность (воспитал более 15 кандидатов и докторов наук) награжден почетной грамотой Московского государственного университета им. М. В. Ломоносова.

# От автора

По мере расширения информатизации современного общества возрастает значение прикладной (вычислительной, компьютерной, инженерной) лингвистики, находящейся на стыке глубоко человеческой гуманитарной науки лингвистики (языковедения), изучающей законы развития и пользования языком, могучим средством мышления и коммуникации, и компьютерных технологий, с помощью которых машине передается все большая часть интеллектуального труда человека.

Настоящая работа является попыткой создать учебное пособие по одной из наиболее актуальных современных лингвистических дисциплин, относящихся как к прикладной лингвистике, так и к общему языкознанию в довольно существенных частях этой науки. Создать такое пособие довольно трудно, поскольку информационные технологии развиваются очень быстро и появляются как новые возможности обработки естественного языка, так и новые проблемы. В то же время довольно интенсивно разрабатываются разные частные аспекты проблемы взаимоотношений языкознания и компьютерных технологий. Однако некоторые принципы компьютерного языкознания — компьютерной лингвистики — можно считать достаточно устоявшимися и образующими основу этой сравнительно новой науки.

В течение многих лет я преподавал вычислительную лингвистику, машинный перевод и другие родственные дисциплины на специальных отделениях филологических факультетов советских и российских вузов. Этот большой опыт преподавания и подготовки специалистов мне пришлось сочетать с моим же опытом долголетнего руководства научными и конструкторскими исследованиями и разработками по машинному переводу, научно-техническому переводу и применению компьютеров в информационном процессе, руководства, в ходе которого приходилось использовать тех самых специалистов, многих из которых я же и готовил. Установилась, таким образом, некоторая обратная связь — наверное, не впервые в истории науки и практики. Результатом такой обратной связи яви-

лись достаточно четкие представления относительно того, чему и как целесообразно учить в данной области, какие теоретические положения и какая методика анализа выдержала проверку практикой и временем, а какие подходы и методы оказались непродуктивными.

Книга написана с целью обобщить итоги и наметить перспективы развития компьютерной лингвистики. Компьютерная лингвистика тесно связана с лингвистическим обеспечением информатики. Для последнего уже установилось английское обозначение lingware. Информатические исследования сегодняшнего дня опираются на некоторые главные лингвистические сущности — морфологию, слово, словосочетание, синтаксис, семантику текста (речи), определить и описать роль которых в современной машинной обработке текстов с помощью компьютера и является целью настоящего учебного пособия. Для удобства студентов и изучающих проблематику вычислительной лингвистики рассмотрены основные результаты научных исследований и конструирования интеллектуальных систем, в частности и особенно машинного перевода, новые результаты в этой науке в России и за рубежом. Я старался выделить главное и по возможности просто и доходчиво довести это главное и новое до сведения специалиста и студента.

Что касается содержания таких наук, как вычислительная лингвистика, структурная и прикладная лингвистика, инженерная лингвистика, алгоритмическая лингвистика и других, предметом и объектом которых также являются анализ и синтез естественно-языковых высказываний, то на этот счет существуют разные точки зрения, и можно заниматься специально проблемами содержания каждой из этих наук в противопоставление другим. Однако я в своем пособии старался избегать проблем классификации и уделял больше внимания описанию существа дела. Читатель сам может отнести тот или иной предмет к области той или иной науки, если захочет это сделать. Название «компьютерная лингвистика» представляется мне наиболее удачным в описании идей, методов, проблем, объектов, результатов исследования и конструиро-

вания лингвистических сущностей с помощью новых компьютерных технологий.

Учебников и пособий по прикладному (в таком смысле) языкознанию в настоящее время немного. В 1996 году вышел, по существу, первый такой учебник в С.-Петербурге. Это безусловно нужная и полезная книга, хорошо освещающая многие актуальные вопросы компьютерной лингвистики (Прикладное языкознание 1996). Однако она построена по совершенно другому принципу, чем настоящее пособие. В этом учебнике рассматриваются разные отдельные проблемы этой науки, от методических и общенаучных до конкретных вопросов обработки естественно-языковых материалов и текстов. Удобство такого расположения в том, что можно читать раздел, посвященный данной конкретной задаче, не отвлекаясь на знакомство с другими проблемами. Весьма полезны также учебники А. В. Зубова и И. И. Зубовой (Зубов, 2000; Зубова, 2001; Зубов и др., 2004). Они также построены по принципам группировки проблем по теоретическим и практическим аспектам. Настоящее же пособие написано по принципу усложнения лингвистических единиц, образующих основы лингвистического обеспечения систем автоматической обработки текстов, — морфема, слово, словосочетание, предложение, текст. Таким образом, названные выше учебники и настоящая книга как бы дополняют друг друга в изложении основных проблем компьютерной лингвистики и научении читателя работе с компьютером, составлению лингвистических алгоритмов и решению других актуальных для современного лингвиста задач. Детальные описания соответствующих алгоритмов и программ можно найти также в капитальных трудах Р. К. Потаповой (Потапова, 2001; Потапова, 2002), в сборниках статей по проблемам компьютерной и прикладной лингвистики.

Настоящая книга предназначена как учебное пособие в первую очередь для студентов и аспирантов отделений и факультетов теоретической, прикладной и компьютерной лингвистики вузов, вообще для студентов, интересующихся данной проблематикой, для специалистов, работающих в соответствующих областях, и для всех, кто интересуется проблемами современной лингвистики и ее приложениями.

# Введение

Компьютерная лингвистика занимается в первую очередь решением задач лингвистического обеспечения информатики. Информатика понимается в широком смысле как наука о закономерностях записи, хранения, переработки, передачи и использования информации с помощью современных технических средств. Поскольку мы живем в компьютерный век, имеется в виду осуществление всех этих процессов в основном с помощью компьютера.

Огромная, если не преобладающая, часть информации существует в виде устных или письменных текстов на естественном языке. Поэтому большое значение имеет обработка естественно-языковой информации. Соответствующая проблематика все больше входит в круг интересов прикладной и теоретической лингвистики, а также прикладной филологии, поскольку последняя есть также наука о толковании текстов и приобретает все большее значение в нашу эпоху развития средств массовой коммуникации и деловой прозы.

Главная цель настоящего исследования — обобщить лингвистические проблемы, попытаться найти лингвистические основы обработки текстов на естественном языке посредством компьютера, не отрываясь от их источника в современном мире — информатики с ее теоретическими, практическими, социальными, дидактическими и другими аспектами, генерирующими «вызовы» лингвистике и филологии.

В первом разделе книги речь пойдет об общих представлениях информатики. Второй раздел посвящен проблемам ввода лингвистической информации в компьютер. Следующий раздел содержит материалы, относящиеся к информатической роли, если можно так выразиться, основных языковых единиц — слова, словоформы, словосочетания, предложения, текста. Четвертый раздел — терминология и терминография, которые объединяются под общим названием «терминоведение». Эта лингвистическая наука, по моему мнению, должна быть отнесена к компьютерной лингвистике, поскольку главное слово естественного языка, которое фигурирует во всех информати-

ческих построениях, — это термин какой-то области человеческого знания. В каком смысле терминология относится к лингвистическим основаниям информатики? В том, что термин определяет как форму, так и семантическое содержание, план содержания текстов, вводимых в информатические системы и подвергаемых лингвистическому анализу с помощью компьютеров. Пятый раздел посвящен общим вопросам моделирования лингвистических сущностей для нужд информатики. Здесь рассматриваются как теоретические установки, так и практические подходы к моделированию. В шестом разделе кратко рассмотрена основная проблематика разработки и использования экспертных систем. И, наконец, седьмой раздел посвящен главной проблеме искусственного интеллекта — машинному переводу. Машинный перевод есть комплексная теоретическая и практическая задача, в которой, как в фокусе, отражаются и концентрируются все проблемы компьютерной лингвистики — от слова до способов анализа содержания и синтеза словоформы, предложения и целого текста. Можно спорить с тем, почему именно машинный перевод есть главная проблема искусственного интеллекта. Однако никто не будет оспаривать того факта, что термины «понимание» или «восприятие» тесно связаны с понятием «перевод»: «понять» — это значит *перевести* на язык смысла.

Прикладная лингвистика, в состав которой входят лингвистические основы информатики и компьютерная лингвистика, тесно связана с лингвистикой теоретической, поскольку прикладная лингвистика имеет также свою собственную теорию. В теоретической лингвистике пока нет ответа на многие кардинальные вопросы владения языком и «препарирования» языка для кибернетических целей. Однако это явление закономерно. Оно показывает, что прикладная лингвистика — наука, активно развивающаяся в настоящее время, и в этом развитии используется не только накопленный многовековой опыт изучения языка, но и поразительные интеллектуальные возможности современной техники.

# *Глава 1*
# Информатика и компьютерная лингвистика

## 1.1. О термине «информатика»

Термин «информатика» начинает устанавливать свое основное значение только теперь. Происходит возвращение значения этого слова в русском языке к его первоначальному значению из французского — informatique — вычислительная техника в широком смысле, software и hardware вместе, т. е. как программное обеспечение, так и аппаратное, вычислительной техники и компьютеров. В настоящее время сюда следует включить также и lingware, т. е. лингвистическое обеспечение в виде специальным образом организованных словарей и алгоритмов анализа и синтеза текстов.

В советской информационной традиции термин «информатика» впервые был применен для обозначения научно-технической информации (точнее было бы сказать, информирования) как науки (Михайлов и др., 1968). Ю. И. Шемакин определяет информатику как науку о наиболее общих закономерностях построения и преобразования информационной модели мира, определяющей роль человека и технических средств в процессах обработки информации в технических, биологических и социальных системах (Шемакин, 1985). Президент Международной академии информатизации (МАИ) И. И. Юзвишин обосновал новую науку — «информациологию» — научную основу построения информационной модели мира (Юзвишин 1995). Изменение содержания понятия «информатика» и появление новых родственных понятий и терминов говорит о том, что содержание данной науки находится в процессе становления, что как это содержание, так и предмет и объекты соседних наук постоянно изменяются и взаимодействуют в своем интенсивном развитии.

В особых отношениях информатика в такой трактовке находится с кибернетикой. Кибернетика изучает наиболее общие законы управления (субъект и объект управления, их взаимоотношения между собой и с окружающей средой). Управление рассматривается как процесс преобразования информации о желаемом и фактическом состояниях объекта в управляющую информацию, дающую возможность достичь желаемого состояния. Процессы управления, таким образом, неотделимы от информационных процессов. Однако управление включает и силовые процессы, которых нет в информатике. Каждый вид информации характеризуется своим содержанием, однако все они подчинены общим закономерностям машинного представления.

Основной задачей информатики является изучение закономерностей, в соответствии с которыми происходит создание, преобразование, хранение, передача и использование информации всех видов, в том числе с применением современных технических средств.

Н. Н. Ефимов и В. С. Фролов определяют информатику в более широком плане, как науку, включающую искусственный интеллект (Ефимов и др., 1991). В таком же широком плане трактуется информатика по отношению к кибернетике:

«Процесс широкого внедрения вычислительной техники иногда называют процессом информатизации общества, а все, что связано с применением ЭВМ, их разработкой, созданием программ для них, называют информатикой... Кибернетика определяется как наука об общих законах управления в живой и неживой природе. На современном этапе управление в многообразных сферах "неживой" природы (в производстве, науке, технике) осуществляется с помощью ЭВМ. В этом смысле информатику можно назвать частью кибернетики. ...

В то же время широкое применение вычислительной техники для научных расчетов, научного прогнозирования, анализа результатов экспериментов ... позволяет сказать, что информатика охватывает другой, более широкий класс проблем и решает более разнообразные задачи по сравнению с кибернетикой в классическом ее понимании.»

Если же отвлечься от самых общих определений, то под информатикой понимают науку, связанную:

1) с разработкой вычислительных машин и систем, с технологией их создания;
2) с разработкой математических моделей естествознания и общественных явлений с целью их строгой формализации;
3) с обработкой данных, созданием численных и логических методов решения задач, сформулированных на этапе построения математической модели;
4) с разработкой алгоритмов решения задач управления, расчета и анализа математических моделей;
5) с программированием алгоритмов, созданием программного обеспечения ЭВМ» (Власов и др., 1988, с. 13).

Среди нетрадиционных точек зрения на информатику отметим взгляд на нее как на составную часть массовой коммуникации. Массовая коммуникация — это «общезначимый современный текст, в создании и распространении которого принимают участие новейшие технические средства и устройства: мощные печатные машины, телевидение, кино, магнитофонная запись, компьютеры и пр. Причем это преимущественно текст серьезного характера, служащий главным образом нуждам общественного управления, связанный с развитием, регулированием и устройством современного массового производства» (Рождественский, 1979, с.163.). Информатика противопоставлена всем другим видам текстов как «вторичный» текст. Тексты информатики создаются специализированными органами информатики — информационными центрами и пр. Интегрируются информационные службы, а также и сами тексты, которые образуют массивы сведений.(См. также: Рождественский, 2003).

Нетрудно видеть, что в таком определении информатики она совпадает с научно-технической информацией, с той, однако, разницей, что научно-техническая информация понимается обычно и как тексты, т. е. результат определенной деятельности, и как сама деятельность в отношении таких текстов, — «научно-техническое информирование», распространение и организация распространения и использования научно-технической информации.

Сам язык есть сочетание техники создания языковых знаков, составляющих индустрию языка, как говорит академик Ю. В. Рождественский: техника устной речи, материалы и орудия письма, книжная печать, все виды информационных технологий. Каждая технология создания языковых знаков имеет свои потенции для раскрытия смыслов, широты или узости возможностей сообщения между людьми, возможности зафиксировать и упорядочить наличную культуру личности, общества и его частей и организаций. Особенно важна языковая технология для фиксации культуры. Так, устная речь может формировать культуру лишь в памяти людей, письменная — хранить культурозначимые тексты и оперировать ими, создавая их смысловую организацию в виде каталогов и словарей; речь на компьютерах — моделировать все знаковые системы. Компьютерные технологии ставят язык в положение, когда он максимально детально передает глубинные и профессиональные смыслы, свойственные неязыковым знаковым системам (системы записи речи, игровые и обучающие системы, расчеты, моделирующие системы, электронные системы сочинения и исполнения музыки, компьютерная анимация, изображение движений тела на экране, создание дизайна костюма, архитектуры, системы автоматического проектирования, машинный перевод, системы информационного поиска — хранилища культурных сведений) (Рождественский, 2003, с. 181).

При всем разнообразии приведенных выше определений информатики и характера решаемых ею задач, а, может быть, именно благодаря этому разнообразию, информатику можно определить как точную науку, базирующуюся на естественном знании и логических законах, но использующую при этом также и гуманитарные, нечеткие знания, имеющие теоретико-множественное описание.

## 1.2. Область возникновения лингвистических проблем информатики

Человек живет в информационной среде. Современная информационная технология включает растущее число автоматизированных информационных систем, средств массовой комму-

никации, телесвязи и радиообмена, систем машинного перевода, диалоговых, вопросно-ответных, экспертных систем, помогающих человеку в принятии ответственных решений. Языковая практика информационного общества, а также современные теоретические исследования оперируют машинными фондами языков, алгоритмами и программами автоматической переработки текстов, моделированием речемыслительной деятельности, созданием автоматов, не только записывающих и регистрирующих, но и «понимающих» человеческую речь и мысль.

Общение с ЭВМ развивается в сторону использования удобного для человека естественного языка, а не специальных языков программирования, доступных узкому кругу специалистов. Если раньше общение с компьютером было уделом узких групп специалистов, то теперь для того чтобы дать возможность широким массам коммуникантов пользоваться новейшей компьютерной техникой, диалог с ЭВМ должен быть достаточно простым. Персональный компьютер становится частью интеллектуального технического окружения человека, подобно телефону, телевизору и пишущей машинке.

Человеческий язык, естественный язык для коммуникации сообществ людей, продолжает оставаться мощным и незаменимым средством передачи информации, поэтому ввод последней в компьютер связан с переработкой сообщений на естественном языке. Важно отметить, что у истоков современной прикладной лингвистики лежит форма, поскольку именно от нее происходит дальнейшее развитие способов обработки естественно-языковой информации. Сложные сущности образуются из элементарных, которые особым способом кодируются в памяти машины. В памяти человека информация записана в виде импульсов. «Физиологическими единицами памяти служат пачки нейронных импульсов, способные циклически повторяться» (Лебедев, 1986, с.106). Каждая пачка — как бы буква универсального нейронного кода. Пачки импульсов возникают друг за другом, образуя кодовые слова.

Слово — это уже следующая, более сложная сущность. Сведения из памяти извлекаются двумя способами: рефлекторно, под влиянием запоминаемых сигналов, например при их опознании или заучивании, и спонтанно, т. е. независимо

от воспринимаемых сигналов, примером чему служат тексты художественных произведений. Вследствие циклического повторения кодовых слов, хранящих разнообразную информацию об образах памяти, в том числе образов слов родного языка, воспроизводство последних подчиняется определенной закономерности. Одной из таких закономерностей является, например, известный в статистике речи закон Ципфа.

На другом конце этой цепочки возрастания сущностей как по абсолютной величине физической протяженности, так и по сложности восприятия, находится текст. Филология есть учение о текстах (Рождественский, 1979, с. 223). Исследование текста включает три ступени: анализ конкретных условий его возникновения, изучение условий вхождения этого текста в данную область культуры, общие исторические закономерности понимания и истолкования текста на фоне развития культуры, прогресса в знаниях и речевом общении. Первые два вопроса относятся к частной филологии, последний — к общей. Если, однако, рассматривать вопрос о происхождении текста от мельчайших сущностей до некоторого целого и о дальнейшем функционировании этого текста, например в автоматизированных системах при общении человека с этими системами, то получается, что прикладная филология, будучи определенно частной, занимается всеми этими тремя видами проблем, в том числе и теми, которые отводятся общей. Прикладная филология не конкурирует с общей, однако практически в ее сферу входит все, что так или иначе связано с порождением, существованием, толкованием, хранением, передачей и использованием текстов в прагматическом смысле. Можно было бы утверждать, что прикладная филология отличается от общей в том, что в задачу первой входит широко понимаемое моделирование существенных интеллектуальных функций порождения и различной обработки разных видов текстов с максимальным отделением при этом от неформализуемых особенностей человеческой языковой деятельности (Рождественский и др., 1987, с. 116).

Поскольку устройство и функционирование языка недоступны прямому наблюдению, изучение языка и продуктов его деятельности — текстов разного рода — осуществляется глав-

ным образом с помощью моделирования. Моделирование целесообразно начинать с самых простых и наиболее доступных наблюдению и моделированию языковых уровней. При этом следует обозначить принципиальное отличие одних моделей от других. Модели объяснительного типа принципиально можно применять, не ожидая от них каких-либо результатов в смысле порождения языкового продукта, близкого к тому, который произвел бы человек. Объясняющая (или объяснительная) модель должна лишь непротиворечиво объяснять действие языка в целом, если он моделируется, или каких-либо частей моделируемого явления, если моделируются эти части. Модели же другого типа, воспроизводящего, обретают ценность тогда, когда производимый ими результат подобен тому, который был бы получен в результате деятельности человека. Например, перевод с одного естественного языка на другой может быть осуществлен как человеком, так и компьютером: в последнем случае должна быть использована некоторая воспроизводящая модель перевода, на основе которой построен реализованный на ЭВМ алгоритм и система программ. Результаты можно сравнивать, хотя они могут достаточно отличаться друг от друга по качеству и другим параметрам. В итоге сравнения могут быть получены объективные данные о сущности и этапах процесса перевода, возможностях их формализации и алгоритмизации.

Понятие об уровнях языка заимствовано у традиционной лингвистики. Особое место занимает уровень текста. Учение о тексте, начало которого относится к 1960-м годам, по мнению Ю. В. Рождественского, занялось описанием текста лингвистическими методами и стало как бы заменять филологию в этом вопросе. Это обстоятельство было вызвано тем, что лингвистика должна была отвечать потребностям практики, в частности информационной практики. Деление на традиционные уровни языка удобно тем, что при моделировании практически на каждом из традиционно выделяемых уровней можно найти достаточное число формальных признаков (формальных с достаточно общей точки зрения), на которых можно базировать модели объяснительного или воспроизводящего свойства. Лингвистические знания, отражаемые обычно в словарях,

грамматиках и других описаниях, ранжированные по формальным признакам, дают достаточную основу для моделирования в различных целях.

Ученые предвидят создание систем, превосходящих интеллект человека не только с точки зрения объема памяти, но и в творческих способностях. При этом велика роль слова как активизатора (человеческой) мысли — но также, возможно, и «мысли» робота. «Вероятно, кодовые слова, взаимодействуя, сами по себе способны обмениваться своими фрагментами (отдельными кодовыми буквами и кодовыми словами), подобно тому, как в детском словотворчестве наблюдается создание новых слов из фрагментов известных. В таких обменах ключ к тайнам творчества, к постижению физиологических механизмов интеллектуальных процессов и к построению искусственного интеллекта, превосходящего многократно человеческие возможности» (Лебедев, 1986, с. 115). Порождение же текстов из слов может строиться по достаточно четко изученным к настоящему времени правилам, которые, в частности, могут быть описаны формальными грамматиками разного вида. В качестве примера можно привести следующую схему порождения текста (рис. 1):

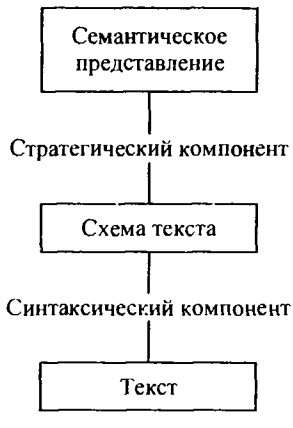


*Рис. 1.* **Схема порождения текста**

Конечно, семантические отношения, на основе которых можно порождать текст по такой схеме, достаточно просты (Aslanides et al. 1993, pp. 27–58). С учетом лингвистических особенностей синтезируемого текста порождение текста может принимать достаточно разработанную форму. Так, в работе Е. С. Андреевой выстраивается цепочка логического порождения текста через ячейку понятийной сети как промежуточного образования между объективным миром и языком, как структурно-логической основы всех языковых отношений (Андреева, 2001).

Построения такого рода рассматриваются обычно в рамках информатики, обращенной к естественному языку. Предметом информатики являются способы и методы представления знаний в системах искусственного интеллекта, информационных системах, моделирующих на различных уровнях человеческую способность к распознаванию, прогнозированию, объединению разрозненных сведений, порождению новых знаний из уже имеющихся. Это все и входит в «информационную модель мира». Коммуникативные потребности общества заставляют по-новому осмыслить язык, связь языка с мышлением. Мышление проявляется через коммуникацию, коммуникация — продукт мышления. Язык рассматривается как система, функционирующая в результате сложного взаимодействия ее составляющих.

Соотношение языка и мышления в коммуникации описано Г. П. Мельниковым следующим образом: «Переход от смысла, т. е. того актуального смысла, который требует знакового выражения, к последовательности знаков речевого потока “включает” механизмы и творческого языкового мышления, и все то, что в процессах связывания смыслов со значениями знаков может быть достигнуто стандартизованными способами, осуществляется языком как автоматическим порождающим механизмом. Следовательно, и до начала коммуникации, и в процессе коммуникации мы имеем дело прежде всего с языком в его статике и динамике, который выполняет роль связующего объекта между актуальными (окказиональными или узуальными) смыслами как единицами мыслительного содержания и языковыми знаками этих еди-

ниц через посредство значений как специализированных коммуникативных абстракций. Поэтому язык в прямом наблюдении, как и другие психические единицы, не дан ни говорящему, ни слушающему. Он является объектом, физически формирующимся в психике, и в этом отношении — идеальным» (Мельников, 1978, с. 277).

Новый состав знаков, новые наборы грамматических значений возникают по законам диахронического развития на основе старых систем знаков. При взаимодействии с ними по правилам диалектики в новой коммуникационной ситуации могут возникать новые знаки. Составной частью новой коммуникативной ситуации являются интеллектуальные сущности, создаваемые человеком и искусственным интеллектом, в виде информационных систем. Новый состав знаков естественного языка, таким образом, возникает в некотором естественно-искусственном многоязычии, которое, по мнению некоторых ученых, создается вследствие распространения в социальном общении информационных языков, взаимодействующих с естественными языками в сферах технологии.

В процессе интеграции информационных систем и включения их в практику языкового пользования в обществе возникают проблемы создания общих языков представления основных видов информации, разработки более совершенных языков доступа к информации и построения языков массового пользования для абонентов. В широком научном контексте филологические и лингвистические проблемы технологического аспекта обработки информации входят в проблематику искусственного интеллекта. Поскольку эти задачи связаны с моделированием речемыслительной деятельности человека в первую очередь в рамках профессиональной деятельности, то для их решения необходимы результаты исследования функционирования языка в процессе обработки профессиональной информации специалистами. Поэтому широко изучается не только современное состояние массовой коммуникации в среде широких потребителей, но и особенности специальной коммуникации, в сфере которой наиболее четко проявляется слияние естественных и искусственных языков (см., например, Рябцева 1986).

В этом отношении интерес представляет дальнейшее развитие способов представления знаний в компьютере. В частности, на основе анализа письменного текста можно выделить в нем составляющие, образующие смысл текста. Здесь в последнее время появились интересные работы отдела прикладного языкознания Института языкознания Академии наук России (Скокан 2002, Новиков 2002). В первую очередь достигнутые результаты применяются к специальной коммуникации, но в то же время они позволяют сделать общие выводы относительно языковой компетенции человека и ее использования. В частности, в том же сборнике в статье Н. К. Рябцевой моделируется естественный интеллект человека. Рассматривается содержание такого понятия, как «представление» и связь этого понятия со знаниями (Рябцева 2002). Подробно изучается и иллюстрируется примерами различие между понятиями *содержание* и *смысл* (Новиков 2002). Восприятие и понимание текста также могут быть смоделированы в целом ряде обучающих программ для компьютеров, причем рассматривается, воспринимается и понимается как письменный, так и устный текст (Потапова, 2002).

Особенности специальной коммуникации заслуживают безусловно больше внимания, чем до сих пор. Показателем увеличения специальной коммуникации является, в частности, перевод научно-технической литературы и документации, в рамках которого проявляются все лингвистические особенности специальной коммуникации. По данным ЮНЕСКО, за 30 лет (с 1948 года по 1977 год) объем переводов разного рода, зарегистрированных в мире, возрос на 600 % (Garnier, 1985). Изменился состав переводов. По данным Европейского Союза, в общем объеме переводов преобладает научный и технический перевод (около 40 %), далее следуют юридический, устный, синхронный, учебный и прочие виды переводов, и, наконец, художественный перевод, с объемом всего 0,3% от общего (Better Translation for Better Communication, 1983). Надо сказать, что такие виды перевода, как юридический и, в некоторой существенной части, политический, имеют те же лингвистические характеристики, что и научно-технический перевод — точное определение терминов, связь тер-

минов с четко очерченными границами предметной области и т. п. (Авербух, 2004).

Эти данные не согласуются с интуитивными представлениями о том, что переводится в современном мире; мы по-прежнему считаем, что переводы Шекспира, Гете и пр. занимают в общем объеме переводов больше места, чем технические и научные описания. На самом деле картина, как видно, обратная.

Состав переводов является хорошим показателем циркулирующей в обществе информации. Скромное место художественных переводов показывает, что языковая практика смещается в сторону деловой, технической и научной «филологии».

Деловая проза занимает особое место в современном языковом общении. «Язык деловой документации, ориентированный на машинную обработку, — это целенаправленно создаваемый технологической сферой вариант языка, отличающийся ярко выраженным аналитическим строем и развитыми свойствами эксплицитного и формального выражения единиц смысла и отношений между ними. Он обладает минимальной по сравнению с естественным языком избыточностью, реализуя известный принцип экономии языковых средств в общении» (Лингвистические вопросы, 1983, с. 25). Участники языкового процесса в сфере управления — а она является одной из наиболее важных социальных сфер — выступают как элементы одной системы. Они знают типовые ситуации, представляющие содержание событий, и общую ситуацию, в которой функционируют системы и происходит обмен сообщениями. Одновременно они знают и возможные изменения типовых ситуаций. В той мере, в какой база данных моделирует реальную систему — фрагмент внешнего мира, — обработка сообщений моделирует ситуации, имеющие место во внешнем мире. В этом случае реализуется свойство естественного языка не только абсорбировать искусственные системы знаков и искусственные средства выражения синтаксических и логических связей между понятиями, но и моделировать внешний мир через его информационное отражение в памяти машины. Это свойство естественного языка могло проявиться полно-

стью только с появлением ЭВМ — универсального преобразователя информации.

Из всего вышесказанного ясно, что компьютер дает не только внешнюю структуру текстов, подлежащих изучению и обработке. Внедрение ЭВМ в информационные процессы затрагивает коренные сущности порождения и использования знаковых ситуаций, моделирование языковых процессов, лежащих в основе мышления и коммуникативного обмена. Например, при анализе текста с точки зрения машинного понимания принимается, что структура текста сводится к логико-композиционному строению. Логико-композиционное строение имеет свою структуру, соответствующую логике предметных отношений. Эту структуру можно в какой-то степени выявить и описать с помощью логических операций над содержанием предложений, составляющих текст. Однако при рассмотрении текста в целом часто оказывается, что его содержание не соответствует логико-композиционной структуре, а строится по своим собственным законам. Целое не складывается автоматически из частей, не представляет собой их арифметическую или логическую сумму. Обращение к тексту в целом, подключение искусственного интеллекта к отдельным частным задачам, например к задаче машинного перевода, позволяет повысить его качество.

Создание естественно-языкового интерфейса, т. е. программных средств, позволяющих пользователю общаться с компьютером на естественном языке, в настоящее время находится в центре работ по искусственному интеллекту (см., например, Диалог 95, 96, 97 или Марчук 1998).

Век «сплошной информатизации», в который нам предстоит войти, будет характеризоваться грандиозной сетью электронной связи с полумиллиардом входов, полной автоматизацией техносферы с парой миллиардов встроенных микроЭВМ, большим количеством (200–300 млн.) персональных ЭВМ и интеллектуальных терминалов, подключенных в сеть связи, десятимиллионной иерархией универсальных ЭВМ, поддерживающих управление обществом и сетью связи, переносом на машинные носители практически всей информации, циркулирующей в обществе. Как будет человек общаться с этой инфо-

сферой, побуждать машины к действию, приобщаться к этому грандиозному фонду знаний, на каком языке общаться с компьютером? Здесь уже есть номенклатура конкретных задач сегодняшней информационной теории и практики: организация диалога с базами данных, составление баз знаний и общение с ними, извлечение смысла из текста в виде команд, фактов, энциклопедических знаний и пр., машинный перевод, синтез текста. Все это общение должно осуществляться на естественном языке (Ершов, 1986). Эти слова Андрея Петровича Ершова, выдающегося советского специалиста по программированию, высказанные почти двадцать лет назад, теперь полностью оправдываются.

Проходящая сейчас в мире инвентаризация языковых данных так или иначе исходит из двух основных тезисов: 1) любые данные о языке могут быть представлены в лексикографической форме; 2) любые лексикографированные данные о языке могут быть переведены в алгоритмизированную, машинную форму (Караулов, 1986). Именно поэтому огромная часть проблематики прикладного языкознания и филологии связана со словом.

Слово является ближайшей дискретной единицей языка, оно имеет точно определенную форму, словоупотребления в тексте разделяются пробелом. В то же время слово — актуализатор человеческой мысли, оно не только и не столько дискретно, сколько вызывает за своей формой некоторый образ. Вообще говоря, можно различать лингвистику аналоговую, непрерывную, и лингвистику дискретную, которую можно представлять с помощью языка цифр. Аналоговая лингвистика — лингвистика образов, эмоций, многие из которых связаны со словом. Современное языковое развитие связано с переходом от аналоговой, непрерывной лингвистики, к лингвистике дискретной (Bertaux, 1984). При этом даже эмотивные компоненты значения слова могут отражаться в словаре, т. е. становиться дискретными (Шаховской, 1986).

Практика прикладной лингвистики обширней, чем теория. Различные практические применения языковедческой науки затронули практически все аспекты владения и пользования языком. Информатика выявила эти аспекты и полученные ре-

зультаты в новом свете; новая информационная технология ставит новые прагматические задачи, ответы на которые сама и находит. О теории прикладной лингвистики и лингвистической теории информатики пока еще можно говорить лишь в том смысле, что эти теории только начинают складываться.

Интересен взгляд на состав того, что мы называем теорией. Габриэль Альтман делает попытку в точных определениях выразить содержание понятий «метод», «процедура», «теория» и др. Теория по Альтману определяется следующим уравнением:

Теория = Концепты, Конвенции, Гипотезы

Концепты составляют необходимое, но не достаточное условие для существования теории. Этот факт часто игнорируется, особенно в гуманитарных науках, где набор концептов часто выдается за теорию. Но теория начинается тогда, когда высказываются гипотезы по поводу отношений между концептами и по поводу процессов, касающихся их.

Конвенции включают в себя определения, операции, правила дедукции, желательные состояния.

Что касается гипотез, то не все предположения относительно дискурса автоматически являются научными гипотезами, которые могли бы быть включены в теорию. Только хорошо сформулированные предположения, которые можно проверить эмпирически, могут считаться научными гипотезами (Altmann, 1993).

Какие именно теоретические импликации прикладной лингвистики видны сейчас? На первое место здесь, видимо, можно поставить лингвистику и логику (Петров, 1986; Петров и др., 1993). Две этих науки, взаимодействуя в течение долгого времени, взаимно обогащались и продолжают продуктивно воздействовать друг на друга. Появились новые виды логики, например иллокутивная логика, отражающая новые взгляды на прагматическую природу текстов. Именно в той мере, в какой лингвистика выходит за пределы высказывания и начинает заниматься текстами в целом, она становится филологией — наукой о толковании текстов. Обработка текстов на естественном языке при общении с ЭВМ вызывает к жизни большое количество теоретических проблем. Например, что такое «пони-

мание»? Что означает то, что машина производит «мыслительные действия»? Может ли она «мыслить»? Современный искусственный интеллект в объеме, требуемом для робототехники, «интеллектуален» в ограниченной степени. Однако если требуется подключить систему искусственного интеллекта к системе машинного перевода для усовершенствования качества последнего, объем знаний, который нужно при этом использовать, становится достаточно велик и может уже быть сравним с интеллектом человека.

Настоящий параграф посвящен основным областям возникновения лингвистических проблем информатики и компьютерной лингвистики. Мы видим, что лингвистические проблемы возникают тогда, когда появляется необходимость измерить алгеброй гармонию почти на всех уровнях языка и текста как продукта языковой деятельности. Сложность и взаимная связь лингвистических явлений разного уровня дают основание рассматривать язык как систему, или систему систем. Язык как система проанализирован Г. П. Мельниковым в упомянутой выше работе. Стройная концепция языковой системы изложена также Владимиром Александровичем Карповым (Карпов, 1992). Языковая система отражает объективно существующий лингвистический универсум, неполнота знаний о котором предопределяет структуру системы. Компьютерная лингвистика позволяет некоторым образом восполнить эту неполноту.

Можно также сказать, что, как и всякая передача информации (коммуникация), словесное сообщение невозможно без материальных средств его передачи и хранения и не тождественно этим средствам (Широков, 2003). Однако материальные носители информации, фактура речи, играют большую роль в существовании, функционировании, совершенствовании и развитии языковой коммуникации (Рождественский, 2003) и тем самым существенно влияют на язык посредством новой организации речи.

## 1.3. Совершенствование массовой и индивидуальной коммуникации

Современное общество характеризуется быстрым развитием средств электронной вычислительной техники, машинизацией производства, развитием транспорта и связи, что сопровождается ростом объема и качества информационных процессов, а именно — сбора, передачи, хранения и переработки информации. Автоматизация, которая началась с простых, рутинных процессов, постепенно переходит ко все более сложным. Развиваются и усложняются автоматизированные информационно-поисковые системы: если ранее многие из них можно было бы назвать автоматизированными библиотеками, то теперь они все чаще становятся составной частью автоматизированных систем управления и систем поддержки принятия решений. Автоматизация информационных процессов означает становление нового типа коммуникации в обществе. Вместе с тем совершенствование массовой и индивидуальной коммуникации связано не только с вычислительной техникой. Вся современная технология, условия существования общества, от увеличения народонаселения до существенных изменений условий быта, коммуникации, перемещения в пространстве и пр. претерпели существенные изменения, которые не могут не отражаться в языке и в практике языкового общения.

Простой пример может проиллюстрировать важность совершенствования языковой практики и массовой коммуникации в современном обществе. Для определения эффективности призывов размещаться в московском метро вдоль платформы, не переполнять хвостовых вагонов поезда, были обследованы пять вариантов текста, передаваемого в метро по радио: *1) Уважаемые пассажиры! Если вы хотите сохранить хорошее настроение, проходите в свободные головные вагоны; 2) Граждане, проходите в головные вагоны, там посадка свободнее; 3) Не толпитесь, товарищи, проходите, пожалуйста, вперед. В первых вагонах более свободно; 4) Граждане, в целях удобного проезда проходите в головные вагоны. Они менее загружены; 5) Не задерживайте посадку, проходите в головные вагоны.*

Тексты произносились одним и тем же диктором и транслировались через динамики, расположенные на всем протяжении платформы, с которой производилась посадка. Результаты экспериментальных наблюдений фиксировались техническими работниками. Расчеты показали, что из пяти использованных текстов три дали положительный эффект (1, 2 и 3), а два — отрицательный (4 и 5). Расчеты по модификации 4 выявили, что она обладает статистически значимым отрицательным эффектом речевого воздействия. Это объясняется тем, что в данной модификации совмещены отрицательные свойства как лексико-семантического, так и структурного планов. Поэтому и возникает так называемый эффект бумеранга — реципиенты совершают действия, обратные тем, к которым их призывают, поскольку текст вызывает у них раздражение и желание поступить наперекор. Эксперимент показал, что при управлении поведением людей немалую роль играет не только содержание текстов, призванных оказывать воздействие на поведение, но и их лингвистическое оформление. Способ лингвистического оформления может не только значительно улучшить эффективность воздействия, но и приводить к результатам, прямо противоположным при несоблюдении правил лингвистического оформления текста (Естественный язык, 1988). Что касается устной речи, то современные исследования в этой области, в том числе и с применением новых технических средств, дают возможность провести, например, довольно тонкие различия между фонетикой и фонологией (Князев, 2001). Лингвистическое оформление, таким образом, позволяет получать разные результаты при вроде бы одинаковом содержании высказывания с точки зрения собственно информации как таковой. А фонетика и фонология позволяют эксплицировать элементы просодии и других средств воздействия в данном конкретном случае.

Говоря о массовой коммуникации в таком контексте, следует отметить, что здесь также на языковое выражение влияют обычаи и узус социальных призывов: так, например, в японском метро объявления типа «Поезд дальше не идет, просьба освободить вагоны» были бы сочтены за грубость.

Взаимоотношения теории и практики в языковом аспекте продолжают оставаться сложными и многоплановыми, однако в части появления новых теорий языка, целого ряда сложных понятий и лингвистических технологий за последние сорокпятьдесят лет можно проследить ведущую роль новых технологических достижений, не связанных напрямую с языком, но повлиявших на теорию последнего. Достаточно назвать идею машинного перевода и всего, что с ней связано. Сама по себе концепция механического перевода возникла достаточно давно, в 1933 году, и уже тогда получила внимание исследователей. Однако вследствие отсутствия технической базы она не была воплощена и никакого реального движения не получила. Только после появления современной вычислительной техники и возникновения современных ЭВМ языковые исследования, стимулированные взглядом на язык как на код, стали развиваться ускоренными темпами.

Можно также вспомнить и другие проекты, повлиявшие на развитие лингвистической проблематики в прикладном плане. Так, известные с давних времен поиски и разработки универсального языка не привели к практическому решению проблемы универсального общения на едином языке, но способствовали развитию интерлингвистики, которая занимается проблемами общих для всех языков свойств и отношений. Здесь можно также заметить, что поиски единого языка занимали большое место в истории машинного перевода, когда некоторые разработчики стремились найти язык-посредник, перевод с которого и на который позволил бы, по их мнению, сократить число алгоритмов. И в настоящее время продолжаются поиски единого языка. В качестве примера можно привести описание эсперанто и его истории, характеризующейся, по мнению некоторых эсперантистов, преследованиями этого языка (Линс, 1999), а также новый международный язык-посредник «эльюнди» (Колегов, 2003). В этих проектах, как и в других, имел место определенный разрыв между теорией и практикой.

Отношение практики и теории в общем плане может быть выражено следующей диаграммой, представленной на рис. 2.

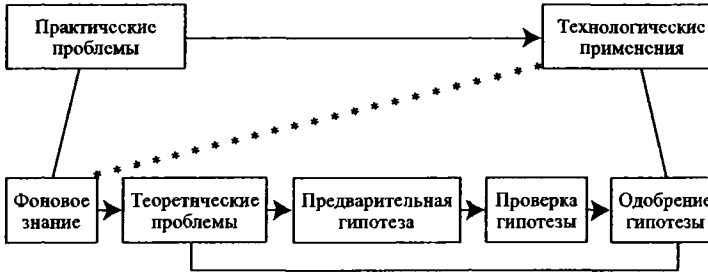


*Рис. 2.* **Соотношение теории и практики (по Кульвейну)**

В данной диаграмме горизонтальная линия (от фонового знания к теоретическим проблемам и далее к гипотезам) показывает рост и накопление научных знаний, а практические проблемы и технологические применения стимулируют этот рост (Kuehlwein, 1987).

Человек как часть некоторой человеко-машинной информационной системы, целью которой является принятие решений, обладает определенными недостатками с точки зрения эффективности, быстроты и правильности принятия решений. Ведь именно для принятия решений и служит информация. Когнитивные ограничения человека связаны со следующими факторами:

1) недостаточная емкость кратковременной оперативной памяти;
2) последовательная обработка данных, медленный процесс записи на хранение данных в долговременной памяти;
3) сокращение и упрощение реального мира при абстрактном рассмотрении его в проблемном пространстве. Эти когнитивные ограничения приводят к срыву когнитивного процесса принятия решения, который происходит в том случае, если информационная потребность для решения проблемы превышает возможности человека в обработке информации. В результате может быть принято неправильное решение (Dubey, 1988). Строго говоря, человеку нужна не только и не столько информация, сколько конкретная информация по данному предмету в данное время.

Специалисты по искусственному интеллекту различают информацию и знания. На простом примере разница в следующем: информационная система, пусть автоматизированная, выдает в ответ на конкретный запрос человека-пользователя названия документов, в которых освещается данная проблема. Однако в подавляющем большинстве случаев человеку нужны не документы с общими сведениями по данной проблеме, а совершенно конкретная информация, которая в выдаче может либо содержаться, либо отсутствовать. Поэтому информация — широкое пространство сведений о данном предмете, а знания — то, что в данный момент нужно пользователю как часть общей информации.

## 1.4. Информатика и компьютерная лингвистика

Информатика как наука и как практическая деятельность, связанная с записью, хранением и переработкой информации, не говоря уже о более широких определениях информатики как науки, занимающейся изучением и построением информационной модели мира, имеет дело как с языком человека, понимаемым как устройство, порождающее устные и письменные тексты, так и самими этими текстами, в которых содержится информация. Разные уровни языка и язык в целом образуют систему. Эта система изучается теорией информатики, связанной с теорией языка. Концепты, которые формулируются в рамках теории и которые в собственном смысле и образуют теорию вместе с наборами гипотез и прагматическими способами проверки теории практикой, допускают объективную проверку, а обратная связь позволяет совершенствовать концепты и разрабатывать более правдоподобные теории.

Лингвистические проблемы информатики, таким образом, начинаются там, где начинается сама информатика. Эти проблемы сконцентрированы в области соприкосновения человека и машины. Основы лингвистических проблем информатики касаются роли языковых и текстовых единиц в понимании, от мельчайших составляющих этих единиц до все более крупных

сущностей — объединений единиц разных уровней. Универсальных лингвистических теорий, объясняющих как язык в целом, так и его отдельные уровни и/или подсистемы, не создано, и неизвестно, возможно ли такие теории создать. Поэтому моделированию и изучению подвергаются отдельные стороны и аспекты языковой деятельности.

Компьютерная лингвистика и является одной такой частью общечеловеческой науки лингвистики (языкознания). Она непосредственно изучает лингвистические основы информатики и все аспекты связи языка, мышления как непосредственной действительности мысли и моделирования этой действительности с помощью компьютерных программ. Компьютерная лингвистика занимается проблемами языковых единиц, меньших чем слово, равных слову, словосочетаний, предложений (высказываний), текста в целом, проблемами моделирования языковых операций, подобных извлечению смысла из текста или перевода текстов с одного языка на другой. Вследствие расширяющихся возможностей лингвистических автоматов растет, расширяется и углубляется проблематика компьютерной лингвистики. В ее рамках разрабатываются и опробируются в эксперименте новые методы и теории.

В настоящем исследовании употребляются термины «прикладная лингвистика», «лингвистические основы информатики», «компьютерная лингвистика». Мы не будем заниматься специальным определением каждой из этих наук. Скажем только, что прикладная лингвистика, по нашему мнению, — наиболее широкий термин из данных обозначений, он включает не только компьютерную и математическую лингвистику, но и лингводидактику и другие науки. Лингвистические основы информатики — наиболее узкий термин из трех. Он в собственном смысле означает совокупность лингвистических проблем, с которыми приходится иметь дело информатике как компьютерной и информационной науке. Наконец, компьютерная лингвистика занимает как бы промежуточное место между этими двумя понятиями. Она шире, чем лингвистические основы информатики, например потому, что может включать лингвистические проблемы обучающих систем, что относится, строго говоря, к лингводидактике. Она у́же, чем при-

кладная лингвистика в целом, поскольку прикладную лингвистику можно понимать как любое приложение лингвистики теоретической, в том числе и безотносительно к компьютерам. По нашему мнению, установившееся теперь понятие компьютерной лингвистики лучше всего подходит к тому материалу, которому мы собираемся уделить основное внимание в настоящей работе.

# *Глава 2*
# Ввод устной и письменной речи в компьютер

Модель человеческой языковой деятельности начинает свою работу тогда, когда осуществлен ввод языковой информации в ЭВМ. Проблемы ввода устного и письменного текста в компьютер всегда занимали значительное место в компьютерной лингвистике. Автоматический ввод предполагает самостоятельное различение машиной знаков текста и отождествление их со знаками языка. То же относится и к устной речи, к звукам человеческого голоса.

## 2.1. Распознавание устной речи. Лингвистические проблемы

### 2.1.1. Автоматическое распознавание звуков устной речи

Автоматическое распознавание звуков устной речи является чрезвычайно актуальной научной и технической проблемой, решение которой позволило бы значительно ослабить влияние факторов, затрудняющих ввод в ЭВМ текстовой информации и управление с помощью каналов связи. Известно, что зрительный канал восприятия информации у человека работает с перегрузкой, в то время как общий объем воспринимаемой им устной информации сравнительно невелик и может быть существенно увеличен. Ввод устной информации рассматривался достаточно давно в связи с развитием таких средств связи, как телефон, однако расширение ввода потребовало новых научных и практических решений, поэтому соответствующие работы в настоящее время развернуты более широким фронтом, поскольку с момента изобретения телефона прошло много времени и коммуникационная технология значительно продвинулась вперед.

Артикуляторный аппарат человека производит звуки, составляющие речь. Каждый звук имеет свои акустические характеристики, которыми занимается физика. Осмысленность звуков в речевом потоке, что и придает речи в целом связность и необходимые семантические характеристики, изучается психологией. Большинство существующих сейчас устройств, распознающих речь, строится на принципе записи акустических сигналов, созданных человеком-диктором, читающим слова. Наборы акустических сигналов, составляющих слова, введены в машинную память. Такой тип распознавания речи имеет три принципиальных ограничения:

— ограничение персональное, поскольку автомат распознает речь только определенного говорящего;
— ограничение языковое — автомат распознает только ограниченное количество слов;
— ограничение в подготовке — автомат распознает речь лишь в тех случаях, когда она заранее подготовлена.

Эти ограничения создают хорошие условия для работы с роботами в автоматизированном управлении и информационном поиске, однако они делают непригодным диалог с системой на естественном языке в массовом обслуживании, например при получении по телефону разного рода справок. Для снятия этих ограничений нужно, чтобы автомат распознавал не отдельные слова, а отдельные звуки и звукотипы.

В основание пофонемного распознавания звуков речи положен анализ речи по длительности и динамике звучания и по чередованию акустического сигнала и пауз, при этом анализируется ритмомелодический контур звучащей речи в связи с паузами (Рождественский и др., 1988). В работе Г. Фанта, Г. Халле и Р. Якобсона давалась универсальная классификация звуков в акустических признаках, причем акустические признаки были поставлены в соответствие с артикуляционными. Критика этих разработок лингвистами сводилась к тому, что акустические признаки в отношении к артикуляционным оказывались недостаточно универсальными. Важнейшим недостатком рассматриваемой классификации является, по-видимому, недостаточный учет слогообразования, акцентуации и ритма, т. е. главных носителей

смысла речи. Отсюда следует, что проблема автоматического распознавания речи недостаточно разработана прежде всего с точки зрения лингвистических представлений об акустике звучащей речи.

Примером связи проблемы распознавания устной речи с лингвистикой является слог. В литературе по лингвистике нет однозначного понимания слога и принципов слогоделения. Л. В. Златоустова выделяет три существенно отличающихся подхода. В соответствии с первым, слог — это единица описания языка, отражающая закономерности сочетаемости фонем. Членение на слоги в рамках данной концепции осуществляется в зависимости от закономерностей сочетаемости фонем, прежде всего от сочетаемости гласного с предшествующим и последующими согласными в конце и начале слова. Второй подход — представление слога как фонетической единицы — связан с тем, что в ряде языков, например в русском, слог непосредственно не связан со смыслом. Для каждого говорящего слог является минимальной произносительной единицей, представляющей собой единство слогообразующего гласного и одного или более согласных. Третий подход — комплексный — учитывает и фонологические, и фонетические категории определения слога. Структура слога определяется правилами фонотактики языка, т. е. как последовательность артикуляционных движений, акустических сигналов, воспринимаемых стимулов (Михайлов и др., 1987).

Изучение слога как комплексной речевой сущности имеет большое значение для распознавания устной речи. Слог является тем фундаментом, на котором строится вся речеязыковая иерархия; с другой стороны, слог выступает как специфическая минимальная единица речепроизводства и восприятия. Функционирование слогового сегмента в потоке речи дает возможность дифференцировать сообщение на дискретные составляющие и одновременно интегрировать последние в целостную информативную в коммуникативном плане структуру. Принцип слогового квантования речи делает возможным сам факт реализации речевого высказывания и его восприятия. Корректно реализованная слоговая последовательность с учетом правил сегментно-супрасегментного характера ведет

к адекватной передаче речевой информации в процессе коммуникации (Потапова, 1989, 1997, 2001).

Особо следует сказать о роли статистики в исследовании речи и идентификации речевых единиц. Обобщающие работы в этом направлении можно найти в обширной современной литературе на данную тему, в частности, в статье Г. Сильницкого (Silnitsky, 2003).

В настоящее время наиболее доступной формой максимально точной фиксации звучащей речи, прежде всего ее тембров и динамики, являются спектрограммы. Поэтому автоматическое пофонемное распознавание речи строится на материале спектрограмм, которым, как предполагают, можно дать фонемную интерпретацию, переводя спектрограммы в фонологическую транскрипцию. При построении автоматического распознавания речи по данным спектрограмм необходимо детализировать и описать: сведения о минимальном наборе распознаваемых единиц, наборы полезных признаков для каждой из распознаваемых единиц, данные о возможных комбинациях распознаваемых единиц в речевом потоке, сведения о существенных для определения исследуемых единиц событиях, содержащихся в спектральных представлениях речевых сигналов, сведения о характере акустических взаимодействий распознаваемых единиц в речевом потоке, правила лингвистически ориентированной сегментации речевого потока с опорой на обнаружение в речевых спектрах лингвистически значимых акустических событий, правила фонетической интерпретации выделенных сегментов в терминах полезных признаков распознаваемых единиц, правила определения слов на основании полученной фонетической информации.

Нужно, таким образом, априорно задать необходимое и достаточное число сегментов на фонемном уровне, описать лингвистические признаки каждого сегмента, затем описать окружение фонемных сегментов, влияющее на их акустическое представление, а также охарактеризовать влияние на представление сегментов нижележащих уровней просодии, зафиксировать взаимовлияние звуков в речевом потоке, дать правила фонологического транскрибирования отдельных фонем, их сочетаний и слов. Эти задачи решаются как путем соз-

дания соответствующего теоретического механизма, так и путем наблюдения за работой автоматического механизма, действующего на основании такого представления.

В компьютерной лингвистике описаны специальные механизмы, обеспечивающие удовлетворительный переход от фонем к аллофонам и обратно. Механизм унификации пригоден для перехода от отдельных фонем к слогам, так как позволяет учитывать признаки в некоторой унифицированной системе их учета и использования в классификации. Так называемая трансдукция, реализуемая в виде автомата с конечным числом состояний, дает возможность переходить от аллофонов к фонемам (Carson 1988).

### 2.1.2. Распознавание изолированных слов

В практике управления роботами важно распознавание изолированных слов. Применяются различные устройства, основанные на таком распознавании. Структурная схема типичного устройства распознавания изолированных слов представлена на рис. 3.

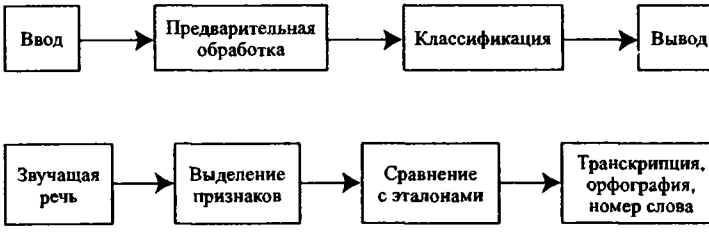


*Рис. 3.* **Структурная схема устройства распознавания изолированно произнесенного слова**

Если слова произносятся четко, то с помощью существующих методов можно использовать словарь, содержащий минимум 100 слов. Приспособление к разным дикторам происходит большей частью путем «доучивания» системы. При пословном распознавании речевой сигнал (слово) записывается через микрофон и подвергается предварительной обработке. Распространена спектральная обработка с помощью ряда полосовых

фильтров. Эта процедура может быть применена в аналоговой форме и в реальном масштабе времени. Общее у всех методов то, что они создают набор признаков, которые в качестве функции времени описывают свойства сигнала и в целом создают «образец» слова (Потапова, 1989, с. 41).

Наконец, в распознавании устной речи есть проблема перехода от фонемы к графеме. Остановимся в этом вопросе на статистическом подходе к нему.

Устройства распознавания устной речи редко используют распознавание слов, только в случае робототехники. В большинстве случаев речевой сигнал препарируется на единицы, меньшие, чем слово (фонемы, слоги и пр.), и затем применяется алгоритм превращения фонем в графемы. В большинстве случаев эти алгоритмы строятся по определенным лингвистическим правилам, по крайней мере, для европейских языков. Однако практика показывает, что эти правила носят статистический характер; правильность и эффективность их применения подчиняются статистическим законам. Поэтому был предложен прямой статистический метод выявления закономерностей конверсии фонем в графемы, основанный на хорошо известных в статистике связях Маркова. Надо сказать, что статистические закономерности довольно часто используются для выявления лингвистических законов и связей и для решения практических вопросов вычислительной лингвистики и информатики. Среди используемых статистических закономерностей цепи Маркова, с помощью которых выявляются закономерности организации текста и лингвистических составляющих текста, занимают одно из видных мест в мировой лингвистической науке. Метод преобразования фонем в графемы с использованием марковского подхода описан в (Rentzepopoulos et al., 1993). В последней работе Р. К. Потаповой марковскому процессу уделено большое внимание. Она пишет, что именно появление техники моделирования скрытым марковским процессом, которая подразумевает стохастическую природу сигнала, повлекло за собой необходимость создания специальных баз данных. Практически все современные системы распознавания речи, обладающие достаточно мощными характеристиками (объем словаря — 1 000 и более слов, многодикторские или обладающие возможностями

быстрой адаптации к диктору), выполнены на основе моделирования речевого сигнала скрытым марковским процессом. Эта техника сейчас безусловно доминирует среди остальных подходов (Потапова, 1997).

## 2.2. Распознавание графем. Исправление искаженных знаков текста

### 2.2.1. Начальная характеристика естественно-языкового текста

Начальной характеристикой естественно-языкового текста, введенного в память ЭВМ, является его буквенный состав, написание, графематика. Буквы алфавита, знаки препинания, небуквенные графемы разного рода — апостроф, кавычки, тире, скобки и пр. — служат в качестве составляющих любого естественно-языкового текста. Они не только не могут быть исключены из него при предредактировании, но и не должны, поскольку являются носителем важной лингвистической и текстологической информации, могущей быть существенной при анализе текста.

Появление идеи машинного перевода, применение к задаче перевода с одного естественного языка на другой электронных вычислительных машин вызвало к жизни большой интерес к графематике, формальному составу символов естественного языка. Было обращено особое внимание на избыточность естественного языка в разнообразных аспектах, в лингвистике появились такие понятия, как энтропия, вероятность, информация, заключенная в символах языка. Возникли первичные представления о составе наук, включающих новые области исследования: математическая лингвистика, статистическая лексикография, теория информации, теория лингвистической реконструкции, комбинаторика, инженерная лингвистика и пр. Определение слова как последовательности символов между двумя соседними пробелами дало возможность применить к изучению форм слова и их распределению в тексте методы математического, теоретико-информационного, дистрибутивно-статистического и других видов точного анализа, возможности

которого были существенно расширены благодаря применению вычислительной техники.

Многочисленные исследования графематического уровня языка с точки зрения теории информации, статистики и теории вероятностей выявили самые различные особенности текстовых структур. Мы назовем здесь лишь некоторые из работ, в том числе и вышедшие достаточно давно, в которых обращается внимание на различные стороны графематического состава слов и текстов с точки зрения точных наук и в преломлении определенных практических задач: (Харпер, 1957; Пиотровский и др., 1977; Яглом, 1957; Wiener, 1949; Oettinger, 1949) и многие другие работы. Уже в первых трудах по кибернетике и по машинному переводу (см.: Винер, Харпер, Яглом, Эттингер и пр.) ставятся актуальные ныне вопросы о буквенном составе современной письменности, о количественных закономерностях, которым подчиняются тексты в графематическом своем составе, о частотности слов, о типах слов и пр. В них исследуются статистические законы комбинации знаков алфавита, вероятности определенных сочетаний букв, частотность слов и словосочетаний и прочие формальные и количественные особенности графемного состава естественных языков. В рамках нашего настоящего исследования мы остановимся на двух аспектах графематики, которые представляются нам в настоящее время наиболее актуальными, а именно: анализ графематического уровня с помощью полиграмм и дериватология.

Из символов любого алфавита можно построить гораздо больше естественно выглядящих слов, чем это есть в любом естественном языке. Комбинаторные возможности буквенных сочетаний заданной длины гораздо больше, чем реально существующие в каждом языке сочетания букв в виде осмысленных слов. Этот факт известен специалистам и по-разному может быть использован в прикладных целях. Например, в Швеции несколько лет тому назад был с помощью компьютера получен список новых имен и фамилий, хорошо звучащих по-шведски, с тем чтобы разбавить существующее однообразие имен. Однако в целом порождение новых слов в естественном языке за счет использования реально существующих комбинаторных возможностей встречается не часто. Интересно, что даже в це-

лях порождения новых товарных знаков комбинаторные возможности сочетаний букв используются достаточно редко (Соболева и др., 1986).

Анализ графематического уровня с помощью полиграмм производится с самыми различными целями, главным образом, для обработки не найденных в словаре слов и исправления искажений. Не найденное в словаре слово может быть: а) отсутствующим в словаре по причине неполноты словаря, хотя оно и является полноправным словом данного естественного языка; б) искаженным словом естественного языка, вследствие чего оно и оказалось не найденным программой поиска по словарю, работающей с введенными в память ЭВМ словами. К числу новых слов, как правило, относятся также имена собственные, географические названия, новые товарные знаки и т. п.

Графематика обладает информационной избыточностью, которая может быть использована для обработки «новых» слов в системах автоматической переработки естественно-языковых текстов, а также для сжатия текстовых данных. В написании слов существуют закономерности, определяемые нормами грамматики.

Для использования этой избыточности в прикладных целях достаточно часто применяются вероятностно-статистические средства. Так, древесно-стохастическое представление графематической информации позволяет решать вопрос о восстановлении искаженных знаков текста. Стохастическое дерево — это дерево, в узлах которого записаны элементы моделируемого объекта, а в дугах — вероятности перехода от предыдущей цепочки узлов к следующему узлу. Применительно к графематике слов в узлах записываются графемы, а в дугах — вероятности следования данной графемы за предыдущей частью полиграммы. Графематика слова имеет неслучайную структуру. Слова построены из определенных полиграмм: состав и количество этих полиграмм определяются рядом факторов, важнейшими из которых являются фонематическая система данного языка, правила орфографии, принятая система обозначения звуков на письме. Несмотря на то что в реальных текстах встречаются далеко не все из теоретически возможных полиграмм, графематика слова не может быть описана анали-

тически и компактно. Для ее описания предполагается вероятностный аппарат (Коростелев, 1985).

### 2.2.2. Диагностика искажений в словах

Диагностика искажений в словах основана на предположении, что среди полиграмм, входящих в искаженное слово, найдется такая, которая не вложится в стохастическое дерево, или ее частота будет весьма низка. Алгоритм диагностики следует этой гипотезе: слово разбивается на полиграммы, и каждая из них вкладывается в дерево. Одновременно с вложением она получает оценку на основании частот, записанных в дугах. Если данной полиграммы в дереве нет или ее частота ниже заданного частотного порога, то в слове фиксируется искажение. Численная оценка правильности слова складывается из оценок составляющих его полиграмм.

Порождение цепочек в стохастическом дереве есть процесс, содержательно не ограниченный длиной полиграмм. Однако генерация должна быть ограниченной, чтобы не быть бесконечной. Л. Ю. Коростелев на материале английского языка для системы машинного перевода АМПАР (Марчук, 1983) выбрал следующие ограничивающие параметры, корректирующие порождение: длина корректирующей цепочки, вероятность очередной графемы корректирующей цепочки, правильность порождаемого прообраза, мера сходства с порождаемым словом. Длина цепочки не может отличаться от длины заменяемой цепочки больше, чем на максимально допустимое количество искажений, определенное в данный момент в автомате. Вероятность очередной графемы корректирующей цепочки выбирается каждый раз максимальной. Правильность порождаемого прообраза проверяется каждый раз диагностирующим автоматом. Мера сходства с восстанавливаемым словом представляет собой алгоритмически несложную процедуру, дающую численное расхождение между двумя цепочками символов (Коростелев и др., 1987).

Параметры корректирующего автомата не закреплены. В начале работы с каждым словом автомат пытается найти варианты прообраза, минимально отличающиеся от искаженного

слова и имеющие максимальное значение правильности. Если диагностирующий автомат обнаружил в слове несколько искажений, обрабатывается наиболее «грубая» ошибка. В случае неудачи — не нашлось прообразов — значения параметров начинают плавно меняться, становиться «мягче» — постепенно увеличивается максимально допустимое количество искажений, снижается порог правильности, позволяется обработка менее грубых ошибок.

Проверка описанных автоматов на ЭВМ показала удовлетворительную их работу. Была определена зависимость количества диграмм, триграмм, тетраграмм и пентаграмм от количества слов, обработанных исследующим автоматом. Результаты представлены на рис. 4 в виде экспериментальных кривых роста. Видно, что эти кривые имеют логарифмический характер, рост количества полиграмм стремится к нулю с ростом количества слов.


*Рис. 4.* **Зависимость количества полиграмм от количества обработанных слов**

Многочисленные тесты по диагностике искажений и их коррекции, проводившиеся также и с помощью генератора псевдослучайных чисел, показывают высокую надежность и работоспособность данных автоматов. Диагностирующий автомат работает с достоверностью не ниже 95 % для триграмм, тетраграмм и пентаграмм, достоверность работы корректи-

рующего автомата колеблется в пределах 75–85 % для этих же полиграмм.

Таким образом, при надежной статистике комбинаций из определенного числа знаков письменного текста можно автоматически проводить достаточно эффективную коррекцию искажений в тексте, которые неизбежно встречаются практически в любой лингвистической задаче прикладного характера.

## 2.3. Лингвистическая дешифровка

Описанный выше подход является одним из методов получения информации из текста. Свойства сочетаний элементов текста позволяют получить из текста гораздо больше информации чисто формальными методами, чем об этом принято думать. Существует целая отрасль лингвистических приемов, объединенных названием лингвистической дешифровки. Приемы лингвистической дешифровки, несмотря на то что в настоящее время исследователи не уделяют им большого внимания, на самом деле представляются весьма перспективными.

В настоящем разделе мы не будем касаться многолетней истории шифровального и дешифровального дела, хотя в целом эта практическая проблематика внесла большой вклад в развитие собственно лингвистических дисциплин и в становление прикладной лингвистики. Частично мы освещаем этот вопрос в разделе о машинном переводе. Однако лингвистическая дешифровка имела место задолго до появления компьютеров и собственно компьютерной криптологии и криптографии, примером тому могут служить древние письменности и рукописи (см., например, D. Kahn. The Codebreakers. Kahn, 1968; также в русской версии Кан, 2000). Мы рассмотрим здесь лингвистические вопросы дешифровки как приема лингвистического описания.

### 2.3.1. Лингвистическая дешифровка как прикладная дисциплина

Б. В. Сухотин считает, что лингвистическая дешифровка есть прикладная дисциплина, которая должна использовать

знания о языке, накапливаемые в других областях языкознания. Критерии цели и качества, в которых нуждаются общелингвистические исследования, можно сформулировать с позиций дешифровки. Критерий такого рода может состоять в том, что можно считать адекватными такие результаты исследования в общей лингвистике, которые обеспечивают возможность дешифровки текстов на произвольных языках. Этот критерий можно назвать дешифровочным принципом (Сухотин 1984).

Решение дешифровочных задач, по мнению Б. В. Сухотина (на наш взгляд, совершенно справедливому), есть естественный путь к созданию конструктивной общей лингвистики.

При таком понимании задач и целей лингвистической дешифровки следует иметь достаточно широкое ее определение. Под лингвистической дешифровкой понимается деятельность, направленная на создание методов распознавания явлений языка в текстах на языках, предполагаемых неизвестными. Эти методы опираются исключительно на свойства сочетаний элементов текста. Главный интерес лингвистической дешифровки состоит в поисках наиболее общих закономерностей строения языка (Сухотин, 1976).

Например, дешифровка незнакомых письменностей базируется на формальных понятиях, с помощью которых строится исходная система знаков, применяемая при раскрытии языковых сущностей из текста. Формальные определения базируются на малом наборе простейших неопределяемых понятий и представляют собой утверждения, построенные по правилам, принятым в математике. Формальный подход позволяет легче и точнее обнаруживать сходство и различие определений, устанавливать их эквивалентность, трансформировать определения в более удобную для конкретной задачи форму при сохранении эквивалентности, обобщать их или сужать по мере надобности. Формальные определения могут быть обобщены и служить целям распознавания при анализе или синтезе соответствующих явлений. Границы их применимости если и не всегда задаются явно, то легко выводятся из формы определений. Формальное определение должно содержать указание на то, как опознать определяемое явление в тексте описанного вида.

### 2.3.2. Статистические методы

Дешифровочный подход требует точного определения роли статистических и теоретико-информационных методов в языкознании. Свойства сочетаний элементов текста проявляются в их статистических характеристиках, индивидуальной и совместной встречаемости. Каковы вообще онтология и гносеология случая в лингвистике? Чисто онтологических категорий не существует, если их понимать как полное и точное отражение реальности (Лесохин и др., 1982). Всякая онтологическая категория есть в определенной степени гносеологическая абстракция, построенная по наблюдениям за объектами и явлениями внешнего мира. М. М. Лесохин, К. Ф. Лукьяненков и Р. Г. Пиотровский считают так: несмотря на то что начальное состояние лингвистического объекта фиксировано достаточно точно и его протекание подчиняется детерминированным законам, случай входит в конечное состояние через ошибки и флуктуации в измерениях либо через возмущения, связанные иногда со взаимной адаптацией экспериментатора и изучаемого материала. Однако полностью случайным процесс порождения речи назвать нельзя, его нельзя уподобить, например, ситуации типа броуновского движения молекул, которое всегда случайно полностью (Herdan, 1956). Порождение текста имеет определенную целенаправленность, которая, хотя и затухает на достаточно большой протяженности текста, всегда действует в нем на коротких дистанциях. Эти особенности действия случая в языке и речи заставляют лингвистов проявлять известную осторожность при использовании классического аппарата теории вероятностей и статистики.

Информационные измерения, опирающиеся на обработку распределения безусловных вероятностей, имеют в языкознании ограниченное применение. Дело в том, что фонемы, слова и другие языковые единицы выступают в тексте в качестве зависимых лингвистических событий, обусловленных контекстом, а их вероятности являются условными. Распределение вероятностей определяется тем положением, которое занимает данная единица в тексте. В большинстве случаев лингвистический опыт характеризуется не безусловной, а условной энтро-

пией, определяющейся тем контактным окружением, в котором находится данный участок текста.

Статистические методы позволяют не только опознать знаки речи на этапе ввода их в компьютер, но и получать гораздо более важную информацию. Так, Мартыненко в работе «Семиотика статистики» выделяет прагматику статистики, семантику статистики, синтаксис статистики (Martynenko, 2003).

Связь с контекстом может служить основой модели текста и особого вида грамматики для извлечения смыслового содержания текста (Descl'es, 1993). К этой концепции мы еще перейдем, а сейчас рассмотрим явления на уровне графем.

### 2.3.3. Графематический уровень

На графематическом уровне проявляются также законы комбинаторики. Лингвистическая комбинаторика — отрасль языкознания, изучающая в рамках лингвистического времени качественные и количественные характеристики как языковых континуумов, так и входящих в них языковых элементов с целью определения возможности (нескольких возможностей или невозможности) и результатов различных видов их взаимодействия.

Речь идет об анализе совместимостей или несовместимостей различных конфигураций данной системы. С точки зрения комбинаторики комбинации полиграмм и других формальных графических элементов в составе словоформы определяются сложными законами плана выражения и плана содержания.

М. М. Маковский пишет: «Приведенный материал дает возможность полагать, что протяженность и состав корня являются стохастическими величинами, всецело зависящими от диалектики плана выражения и плана содержания. С одной стороны, до тех пор, пока комбинаторное разнообразие фономорфологических сегментов в пределах словоформы ограничено определенным семантическим континуумом, а комбинаторная схема не допускает дальнейшего развертывания этого континуума, лексическая таксономия не нарушается.

С другой стороны, до тех пор пока комбинаторное разнообразие элементов той или иной семантической последова-

тельности ограничено определенным вариативным континуумом одной и той же словоформы, а комбинаторная схема не допускает дальнейших комбинаторных изменений указанной слоформы, адекватность этой семантической последовательности сохраняется. Вместе с тем одна и та же словоформа, сочетающаяся с различными семантическими последовательностями, обнаруживает неодинаковую комбинаторику и вариативность фономорфологических сегментов, а одна и та же семантическая последовательность, сочетающаяся с разными словоформами, развертывается по-разному» (Маковский, 1988, цит. с. 29). Из этого текста видно, что сочетания графем можно изучать не только чисто статистическими методами, не проникая в лингвистическую сущность буквосочетаний, но и пытаясь, на основе осмысленности, выявить смысловые ограничения и вообще зависимости комбинаторики графем.

Мы рассматриваем все эти позиции и теоретические положения графематики, главным образом, с точки зрения их практических применений в целях информатики. Это значит, что имеется направленность на описание тех тенденций, которые напрямую связаны с общением человека и компьютера. Речь как комбинация знаков на графемном уровне позволяет строить алгоритмы, распознающие слова, выделяющие новое слово и отождествляющие искажения в отличие от новых слов. Статистико- и теоретико-информационные свойства слов как совокупности графем позволяют, как было сказано выше, применять дешифровочные методы, позволяющие, в частности, отличать текст на естественном языке от нетекста. Таким образом, графематический уровень служит основой для применения целой отрасли науки — дешифровочной лингвистики, характеризующейся специальным подходом, теорией и совокупностью приемов. Значение этой науки в настоящее время возрастает в связи с тем, что получают все большее распространение методы защиты информации от несанкционированного доступа, а криптография и криптология существенно связаны с лингвистическими знаниями и представлениями (Kahn, 1968).

### 2.3.4. Дериватология

Еще один актуальный вопрос для информатики в лингвистическом аспекте на уровне графем: дериватология (морфемография). Она зародилась в недрах лексикографии. Составители словарей все чаще приходят к мысли о том, что нужно включать в словари сведения о словообразующих морфемах. Приводимые в словарях сведения о производных формах состояли из словообразовательного значения, морфологических вариантов, источника заимствования; иногда указывались синонимичные аффиксы, определялась стилистическая принадлежность аффикса, в том числе его употребимость в определенной терминологической сфере.

Усилиями большого числа исследователей заложены основы морфемографии (в частности, дериватологии) английского, немецкого и русского языков. В. И. Бартков, которого, видимо, можно считать основоположником этой науки в нашей стране, обозначает потребность в следующих типах морфематических — дериватологических — словарей:

— прямых и обратных. Префиксальные, циркумфиксальные и корневые морфемы целесообразнее приводить в прямых морфемариях, а суффиксальные — в обратных;

— морфемариях продуктивности, которые содержат информацию о том, сколько разных слов с данной морфемой представлено в языке;

— частотных, в которых приводится информация о том, сколько раз встречается в тексте определенной длины данная морфема;

— морфемариях функциональных стилей, содержащих сведения о том, где употребляются слова с данной морфемой;

— морфемариях идеографических (тезаурусных), в которых представлены морфемы, сгруппированные по их значениям (семам);

— морфемариях, содержащих сведения об этимологии, времени возникновения моделей, деривационной валентности, фонемном и морфемном тяготении формантов и пр. (Бартков, 1982).

Для английского языка, по данным В. И. Барткова, проанализированы все суффиксальные существительные и прилагательные, содержащиеся в словаре Concise Oxford Dictionary (1959) — около 50 тыс. слов. Каждый суффикс охарактеризован по 8 параметрам, среди которых количественные (диахроническая продуктивность, членимость, частеречная принадлежность производящих основ, фонемное тяготение суффиксов к основам, акцентно-ритмическая структура, морфологические варианты). В словаре словообразовательных элементов немецкого языка (1979, под ред. М. Д. Степановой) описание каждой дериватемы дано в виде словарной статьи, содержащей все сведения по 8 параметрам (статус, констатация факта продуктивности и частотности, происхождение, род деривата и пр.); всего описано около 770 аффиксов, полуаффиксов и компонентов сложных слов, давших длинные ряды производных. Этот дериватарий является в настоящее время единственным по числу рассмотренных аффиксов и полноте описания их семантики. Для русского языка нет таких подробных данных, однако основная информация о 87 префиксах и 530 суффиксах содержится в Русской грамматике Н. Ю. Шведовой и в грамматике И. Г. Милославского, причем в последней особо подчеркивается возможность прикладного ее использования, поскольку описание сделано как от значения к форме, так и от формы к значению (Русская грамматика, 1980; Милославский, 1987).

Содержательное описание дериватологических элементов в виде словарей или списков морфем позволяет перейти к изучению статистических закономерностей их распределения, если это необходимо и если это дает какие-либо дополнительные сведения для той или иной конкретной языковой или информатической задачи. В. И. Бартков специально рассматривает применение корреляционных методов к дериватологии, приводя оценку многих из них по результативности в приложениях к конкретному материалу (Бартков, 1984).

## 2.4. Технологии обработки естественного языка в науке и промышленности

### 2.4.1. Ввод речи (текста) в компьютер

Рассмотренные выше подходы к распознаванию устной и письменной речи на уровне фонем и графем дают возможность осуществлять ввод информации в компьютер и производить дальнейшие операции над текстом в самых различных целях. Остановимся на некоторых технологиях распознавания, причем начнем с устной речи.

Современные системы распознавания речи включают различные уровни, каждый из которых несет свою функциональную нагрузку: акустический, параметрический, лексический, синтаксический, семантический и прагматический (Потапова, 1992). Целью современных систем распознавания речи является использование как можно больше неакустической информации, особенно информации более высоких уровней, т. е. семантической и прагматической.

Для целей эффективного использования лингвистической информации на входе должны быть только те предложения, которые описывают определенную ограниченную предметную область.

Желательно, чтобы на входе системы распознавания речи была слитная речь любого пользователя. Однако большинство систем распознавания слитной речи в настоящее время создано для ограниченного числа пользователей. Распознавание речи, основанное на анализе акустического сигнала, требует подробной акустической характеристики сигнала и тем не менее не гарантирует надежной идентификации фонем. Идентификация конкретных слов требует выделения внутри обобщенных классов дополнительных фонетических подклассов, так что в окончательном виде иерархия фонетических классов имеет форму бинарного дерева решений. Исходя из имеющейся обобщенной классификации предлагается процедура построения оптимального дерева решений.

На первом этапе строятся терминальные цепочки — минимальные фонетические классы, необходимые для идентифи-

кации всех слов словаря. При выборе терминальных цепочек используются следующие критерии: множеству фонетических противопоставлений должны соответствовать максимально простые акустические средства; результатом должно быть минимальное число фонетических противопоставлений.

На втором этапе формирования дерева решений производится объединение двух классов низкого уровня (начиная с терминальных) в один класс более высокого уровня, и так до уровня обобщенных классов. В качестве критерия при объединении используется критерий минимального сокращения числа обобщенных классов, содержащих по одному слову.

Имеется достаточно много систем, реализующих эти принципы для распознавания устной речи в прикладных целях.

Изучение системы речевой коммуникации отражает интересы электросвязи, лингвистики, физиологии, психологии и информатики, а также других научных дисциплин. Строение системы речевой коммуникации как большой системы отличается значительной сложностью, характеризуемой многоуровневой иерархической структурой, скрытыми для наблюдения внутренними информационными процессами, диффузностью функциональных признаков. Существенную особенность системы представляет ее биологический аспект, порождающий свойства самоорганизации и адаптации.

Совместное функционирование подсистем, составляющих большую систему, создает свойство эмержентности (целостности), неизвестное для отдельных подсистем. Научный анализ большой системы должен осуществляться на принципах выявления наиболее существенных свойств системы в целом, не поддающихся непосредственному наблюдению, установления сходства и различия этих свойств, а также сопоставимости известных фактов и явлений при условии достаточной полноты и достоверности приводимых данных (Михайлов, 1995).

### 2.4.2. Человеко-компьютерное взаимодействие

Поскольку человеко-компьютерное взаимодействие начинается с ввода устной или письменной речи в ЭВМ, интересно отметить, что уже на этом этапе содержание и эффективность

ввода и распознавания основаны на том, какая цель и какие задачи преследуются самой системой. В обширной монографии на эту тему В. Н. Агеева и Г. Я. Узилевского (Агеев и др., 1995) описываются четыре подхода к исследованию человеко-компьютерного взаимодействия и диалога, в частности:

- эмпирический;
- антропоморфический;
- когнитивный;
- прогнозно-моделирующий или инженерный.

Эмпирический подход полезен конкретными результатами исследований, без которых трудно оценить методы, способы и приемы взаимодействия и общения пользователя с ЭВМ.

Антропоморфический подход интересен тем, что познание общения «человек-человек» может во многом пролить свет на взаимодействие с таким непростым партнером, как компьютер и программы.

Полезность когнитивного подхода заключается в познании того, как пользователь воспринимает задачи и проблемы с помощью компьютера и каким образом последний может оказать ему поддержку в их решении.

Прогнозно-моделирующий или инженерный подход состоит в разработке средств, с помощью которых можно было бы предсказать, какие методы, способы и приемы взаимодействия и общения являются наилучшими в том или ином случае.

Диалог человека и ЭВМ можно также разделять на две составные части: наружный диалог и внутренний диалог.

Прежде чем приступить к их анализу, рассмотрим некоторые составляющие диалога человека и компьютера. Для осуществления простейшего шага диалога необходимы три составляющие: язык ввода информации (командный язык, планшетное меню, язык функциональных клавиш и пр.); экран, информирующий пользователя о том, как следует вводить информацию и предоставляющий последнему различные формы вывода информации (экранное меню, «окна» и др.); функциональный механизм подтверждения синтаксической правильности введенной пользователем информации. Отметим, что этим механизмом осуществляется обратная связь относительно реа-

лизации интеракции: она имеет место при вводе или выводе информации, подтверждении или отрицании синтаксической правильности введенной информации.

Внешний диалог обеспечивает взаимодействие конечного пользователя с человеко-компьютерной системой в части продуцирования интеракции и именуется пользовательским интерфейсом.

Внутренний диалог обрабатывает лексические единицы, поступающие из внешнего диалога, в соответствии с заданным синтаксисом и преобразовывает их в соответствующие команды и операнды. Последние выполняются специальными прикладными программами. В результате взаимодействия диалогов разрабатывается наиболее оптимальная конструкция действующей системы, реализующей восприятие человеческой речи.

Таким образом, в настоящем разделе мы показали, что уже на первом этапе восприятие человеческой речи в ее устной или письменной форме определяется возможностями и целью всей системы человеко-компьютерного взаимодействия. Структура системы и проблематика прикладной лингвистики далее определяются существом и характером обработки лингвистических единиц большего объема, а именно — слов, словосочетаний, синтаксических структур, связного текста в целом.

# *Глава 3*
# Обработка лингвистической информации на уровне словоформ, слов, словосочетаний, предложений, текста

После того как мы рассмотрели общий характер лингвистических проблем информатики, их состав, происхождение, структуру и перспективы, становится ясно, что конкретные проблемы зависят от типа лингвистической единицы, подаваемой на вход ЭВМ. В нашем рассмотрении мы будем придерживаться принципа следования от наименьшей языковой единицы к наибольшей: от фонемы, графемы, морфемы к словоформе, слову, словосочетанию, высказыванию, предложению, тексту в целом.

Лингвистические проблемы самых мелких, неделимых без потери лингвистической целостности единиц — фонем, графем, морфем, — были рассмотрены в предыдущем разделе. В настоящем разделе будут рассмотрены лингвистические аспекты информатического слова, словоформы, словосочетания, предложения, текста. Наиболее важной частью этой проблематики является морфологический анализ.

## 3.1. Машинная морфология

В данном разделе мы рассмотрим вопросы автоматического морфологического анализа естественно-языкового текста.

Сразу же следует сделать одно существенное замечание. Если в традиционном языкознании «для человека» под морфологией слова справедливо понимается то и только то, что относится к его форме, — окончания, суффиксы, аффиксы, флексии и пр., деление на корень и другие части словоформы, то в автоматической обработке текста на естественном языке

морфологический анализ означает процедуру, в результате которой из формы, внешнего оформления слова в тексте можно получить сведения о самых различных уровнях языковой структуры. Понятие «морфологический анализ» в таком смысле родилось в машинном переводе (Василевский и др., 1970).

### 3.1.1. Автоматический морфологический анализ

Автоматический морфологический анализ языка синтетического типа, т. е. с богатой морфологией и различными формами, начинается с идентификации словоформ и по возможности с группировки их по каким-то функциональным классам. При этом важным моментом является то, что различие между словообразованием и словоизменением, которое играет большую роль в традиционной трактовке морфологического уровня языка в языкознании, в компьютерной лингвистике не играет особой роли (кроме случаев, когда такое различие учитывается в прикладной задаче), и поэтому разница между словообразованием и словоизменением в большинстве случаев стирается.

Так, морфологический анализ русской фразы (например, как первый этап МП с русского на эстонский язык, см. работу Р. Пальма (Пальм, 1962)) состоит из следующих частей:

1) лексическая обработка фразы. Сюда входит использование словаря основ путем выделения исходных слов, распределения их на статьи (лексемы) по морфологическим и семантическим принципам (например, существительные на *-ние* входят в одну статью с формами соответствующего глагола), выделение основ лексем, множество которых и образует словарь;
2) идентификация окончания. Морфологическая информация к основе включает тип основы (частицы, существительные, прилагательные, глаголы), номер таблицы окончаний, морфологическая «сеть» — перечисление форм, имеющихся у данной основы, морфологические признаки — дается перечень признаков у разных типов основ. Вводится понятие «шкалы слова» — перечня возможных грамматических интерпретаций для каждой переводимой словоформы.

В результате предварительного анализа структуры фразы последняя делится на части. Дается классификация предложений на 6 типов в зависимости от их места во фразе и встречаемости в них определенных типов слов. Особым этапом производится обработка эквивалентных форм, а именно: преобразование неоднозначных шкал слов в однозначные = разрешение дизъюнкций информаций, т. е. снятие грамматической омонимии.

Из этого видно, что на самом деле в этап морфологического анализа входит большое количество операций, с помощью которых получается информация, не всегда относящаяся к собственно морфологической. Вследствие этого целесообразно утверждать, что в компьютерной лингвистике понятие морфологического анализа является понятием операционным. Если в традиционной лингвистике к морфологическому анализу относится то, что характеризует форму и отвечает на вопрос «что классифицируется?», то в вычислительной (прикладной) лингвистике важно не «что», а «как» получается та или иная информация, т. е. из формы слова в тексте.

В первые годы работ по машинному переводу было предложено большое количество разнообразного рода алгоритмов автоматического морфологического анализа для языков самого разнообразного строения, отличающихся друг от друга «морфологией». На сегодняшний день задача морфологического анализа — наиболее сложная процедура на уровне слов — может считаться практически решенной, поскольку есть достаточное количество удовлетворительно работающих алгоритмов. Авторы монографии «Лингвистические вопросы алгоритмической обработки сообщений» (Лингвистические вопросы 1983) считают, что за два десятилетия создано по крайней мере несколько десятков алгоритмов морфологического анализа для разных языков, в том числе 10–12 для русского.

В разработке морфологического анализа выделилось несколько направлений. Одно из них моделирует классическую схему анализа путем разделения словоформы на основу и предположительное окончание с последующей проверкой на совместимость окончания с остающейся основой. Другое направление использует информацию, содержащуюся в конечных буквосочетаниях. Эта информация получается в резуль-

тате предварительной статистической обработки словаря. Третье направление развивается в последние годы. На этом направлении создаются универсальные математические модели морфологии в форме открытых систем уравнений, позволяющих путем вычисления осуществлять нормализацию словоформ, получение грамматической информации и синтез словоформ.

В основу построения алгоритмов морфологического анализа положено разбиение всех слов на классы, определяющие характер изменения буквенного состава форм слова. Эти классы могут быть названы морфологическими (Белоногов и др. 1979). Изменения форм слов могут носить различный характер. Они могут быть связаны как с изменением основы слова, так и с изменениями его окончания. Изменение буквенного состава основ имеет место, например, в следующих парах: *сижу — сидишь, шел- шли, тренировка — тренировок, нес — несли, кто — кого, судно — суда, человек — люди.* Изменение окончаний является основным способом образования различных форм слов. В русском языке, например, оно используется как самостоятельно, так и в сочетании с изменением основ слов.

Морфологические классы слов делятся на два вида: 1) основоизменительные классы, характеризующие систему изменения основ; 2) флективные классы слов. Они выделялись, например, для русского языка в системе машинного перевода АМПАР, на основе анализа их синтаксических функций и систем падежных, личных и родовых окончаний. Классы неизменяемых слов выделялись только по синтаксическому принципу. По своей синтаксической функции изменяемые слова объединены в следующие группы: 1) существительные; 2) прилагательные; 3) глаголы в личной форме; 4) глаголы прошедшего времени, краткие прилагательные и причастия; 5) количественные числительные (Марчук, 1983). Флективный класс может быть охарактеризован либо некоторой системой признаков, либо словом-представителем, которое является носителем этих признаков. Признаками, по которым изменяемое слово может быть отнесено к определенному классу, являются: 1) принадлежность к одной из синтаксических групп (или подгрупп); 2) система окончаний (тип словоизменения).

### 3.1.2. Виды автоматического морфологического анализа

Различают следующие виды морфологического анализа:

- морфологический анализ со словарем основ;
- морфологический анализ со словарем словоформ;
- морфологический анализ методом логического умножения;
- морфологический анализ без словаря, с помощью таблиц.

Наиболее распространенным видом автоматического морфологического анализа является анализ со словарем основ, используемый для большинства европейских языков. В этом виде анализа используется словарь основ слов и ряд вспомогательных таблиц. В словарь включены основы простых и сложных слов без внутренней флексии. Если слово имеет несколько форм основ, то в словарь, как правило, включены все формы основ слов. Каждой основе словаря ставится в соответствие сочетание кода основоизменительного класса и кода флективного класса, а омонимичной основе — серия сочетаний таких кодов. Так устроен словарь в системе, описываемой Г. Г. Белоноговым.

Морфологический анализ и синтез слов производятся с помощью словаря основ и ряда вспомогательных таблиц. В словарь включены основы простых и сложных слов без внутренней флексии. Для сложных слов с внутренней флексией типа «*слесарь-инструментальщик*» в словаре приведены лишь основы простых слов, входящих в состав этих сложных слов. Если слово имеет несколько форм основ, то в словарь, как правило, включаются все формы основ слов. Каждой основе словаря ставится в соответствие сочетание кода основоизменительного класса, а омонимичной основе — серия сочетаний таких кодов.

Морфологический анализ слова начинается с его флективного анализа. Последний производится с целью правильного выделения его основы, замены буквенного состава основы ее порядковым номером по словарю и определения грамматической информации слова. Алгоритм морфологического анализа состоит из 32 блоков и учитывает все шаги морфологического анализа с помощью словаря основ, возможные варианты ана-

лиза при отклонении процесса от однозначных правил, переход к следующим ступеням анализа.

Морфологический анализ со словарем словоформ также довольно распространен. Из общих соображений он применяется, когда морфология данного языка достаточно бедна. Кроме того, на первый взгляд представляется, что алгоритм анализа со словарем словоформ проще, чем алгоритм работы со словарем основ: не надо осуществлять членение входной словоформы на морфемы с последовательным поиском по словарю и пр. Однако на самом деле при анализе со словарем словоформ остаются следующие проблемы:

- анализ не найденных в словаре слов. Определение некоторой информации для слова, не обнаруженного в словаре, является необходимым для последующего анализа: например если мы не нашли данного слова, то по крайней мере должны определить его часть речи, чтобы не исключить возможности дальнейшего грамматического (синтаксического) анализа;
- отождествление разных словоформ одного и того же слова. Если каждая словоформа будет выступать как самостоятельная лексическая единица, то это существенно затруднит весь последующий анализ и синтез. Словоформы одного слова должны быть обозначены как таковые. Это означает, что система морфологического анализа со словарем словоформ должна иметь список аффиксов, корней (основ) слов и другие необходимые атрибуты для идентификации разных словоформ одной и той же лексической единицы.

Эти требования фактически сводят на нет преимущества анализа со словарем словоформ, и поэтому анализ со словарем основ применяется значительно чаще.

Особое положение занимает способ автоматического морфологического анализа методом логического умножения. С. Я. Фитиаловым положены начала формальной морфологии (Фитиалов, 1961). Функция, определенная на словоформах и сопоставляющая каждой словоформе некоторую информацию, называется словарной функцией. Всегда имеется возможность задать значения словарной функции в виде таблицы — словаря

словоформ. Однако существуют более экономичные способы задания этой функции. Так, ее можно представить в виде следующей последовательности четырех операций: 1) словоформа как цепочка букв членится на морфемные сегменты; 2) словоформа как цепочка морфемных сегментов заменяется неупорядоченным множеством новых элементов — морфем; 3) словоформе как множеству морфем приписывается некоторая информация; 4) эта информация преобразуется в требуемую окончательную информацию о словоформе.

Каждой морфеме можно сопоставить информацию, получаемую в результате объединения информаций о словоформах, в которые входит данная морфема. Такого рода объединение информаций соответствует дизъюнкции в логической интерпретации. Информация о словоформе получается как пересечение, или логическая конъюнкция, информаций о морфемах, входящих в данную словоформу. Тем самым функция, определенная на морфемах-множествах, заменяется функцией, определенной на морфемах-элементах.

Морфологический анализ методом логического умножения применяется к флективным языкам и предусматривает наличие словаря основ. Сущность метода и применение его к конкретному языку можно видеть на примере алгоритма анализа русских словоформ, предложенного венгерским специалистом Д. Варгой (Варга, 1964). Сначала производится поиск слова в словаре основ. Если слова, имеющие окончания, не находятся в словаре, тогда от каждого такого слова отбрасывается по одной букве справа и поиск повторяется. При отрицательном ответе отбрасывается следующая буква и т. д. Отброшенные буквы образуют окончание и фиксируются. Каждая отброшенная буква считается элементарной единицей морфологического анализа. Ей приписывается булевый вектор — совокупность нулей и единиц, компонентов этого вектора. Число компонентов этого вектора равно числу грамматических категорий, которые могут быть выражены окончанием, частью которого является данная буква. Поскольку предварительно был произведен поиск по словарю основ и установлена часть речи анализируемого слова, имеется возможность одинаковым буквам, входящим в окончания разных частей речи (например, буква *м*

в окончании существительного и прилагательного) приписывать разные векторы.

Пусть, например, требуется определить, в каком числе и падеже стоит существительное *столом*. После поиска в словаре устанавливается, что основа *стол* — существительное, буквы, входящие в состав окончания, *о* и *м*. Буква *м* встречается среди букв окончаний существительного в творительном падеже единственного числа мужского и среднего рода, а также в дательном и творительном падежах множественного числа всех трех родов. Приписываем букве *м* такой булевый вектор, в котором на месте компонентов, соответствующих падежам, в которых она встречается, стоят единицы, а на месте других компонентов — нули. Таким же образом поступаем и с другой буквой окончания. Произведя логическое умножение векторов букв *о* и *м*, получим в результирующем векторе единицу на месте разряда той грамматической категории, в окончании которой встречается одновременно и буква *о*, и буква *м*, а именно в разряде, соответствующем творительному падежу единственного числа.

Морфологический анализ без словаря, или так называемый «независимый» анализ, производится без обращения к словарю, только за счет использования таблиц аффиксов и особого списка не имеющих грамматического значения слов. Такой способ используется достаточно редко.

Каково же современное положение с автоматическим морфологическим анализом?

### 3.1.3. Современное состояние морфологического анализа

Современное положение характеризуется тем, что сильно увеличились требования к качественным показателям систем автоматической переработки текста. Теперь задача создания быстродействующего алгоритма морфологического анализа ставится следующим образом:

- основу системы должен составлять мощный политематический словарь, обеспечивающий покрытие текстов по любой тематике не менее чем на 98–99 %;

- алгоритм анализа должен быть словоизменительным, что позволит при одном и том же объеме распознавать примерно в 8 раз больше словоформ (для русского языка), чем количество лексических единиц в словаре;
- «новые» слова должны обрабатываться наряду со словами, содержащимися в словаре. При этом объем информации для новых слов должен быть таким же, как и для словарных единиц, а вероятность их правильного определения не менее 0,9–0,95 %;
- скорость обработки текстов должна быть, при прочих равных условиях, по крайней мере на порядок выше, чем у существующих процедур;
- на объем исходного текста не должно накладываться никаких ограничений;
- система должна сохранять свою работоспособность в условиях дефицита ресурсов ЭВМ;
- система должна быть обучаемой, т. е. должна иметь средства для пополнения имеющихся словарей и настройки их на различные предметные области;
- процедурная часть системы должна достаточно легко приспосабливаться к меняющимся ресурсам ЭВМ с целью их наиболее оптимального использования, а также иметь возможность работы с различными входными и выходными форматами;
- следует иметь синтаксические средства контроля и корректировки грамматической информации к «новым словам» с учетом микроконтекста;
- должны быть разработаны специальные информационные структуры для представления данных и методы доступа к ним более эффективные, чем структуры и методы, входящие в состав операционных систем ЭВМ.

Массивы таким образом подобранных данных имеют по состоянию на сегодняшний день следующие измерения: политематический словарь словоизменительных основ слов содержит более 100 000 лексических единиц и обеспечивает очень высокое покрытие научно-технической лексики практически любой тематики. Этот словарь был создан в результате обра-

ботки текстов объемом свыше 30 млн. слов. Словарь словоформ, составленный по текстам, включает 46 тыс. лексических единиц, он составлен по текстам объемом более 3 млн. слов. Два этих словаря имеют тщательно выверенные наборы грамматической информации, дающие детальное представление о морфологической структуре слов и их синтаксических свойствах (Зеленков, 1988).

Элементы морфологического анализа довольно сильно выражены даже в языках с грамматическим строем, существенно отличающимися от строя европейских языков. Машинный перевод с китайского языка предусматривает, например, процедуру анализа односложных и двусложных китайских слов на уровне, близком к уровню морфем (Зелко, 1991).

Автоматический морфологический анализ вызвал к жизни специальный тип словарей. Лексические единицы языка упорядочиваются в соответствии с формой и правилами порождения и образования словоформ, по словоизменительным и словообразовательным классам. Одним из видов словарей такого типа, т. е. специально учитывающих требования морфологического анализа, являются обратные словари, применение которых началось от специальных требований лингвистической дешифровки и которые в настоящее время широко применяются в машинном переводе для определения грамматических характеристик не найденных в словаре слов, при анализе словоформ флективных языков (Штиндлова, 1966; Белоногов, 1971; Козьмина, 1988 и пр.)

Особенностью обратных словарей является представление слов словника: сначала идут слова, оканчивающиеся на первую букву алфавита, затем на вторую и т. д. При совпадении последних букв учитываются предпоследние буквы, далее — третьи от конца и т. д. Таким образом, слова расположены в алфавитном порядке, начиная от конца слова. При этом, естественно, объединяются слова, относящиеся к единому словообразовательному или словоизменительному типу, сложные слова с одинаковой последней составляющей.

Обратные словари могут решать достаточно широкий круг задач. Они наглядно представляют морфологические характеристики данного языка. Если грамматические описания часто

содержат утверждения о том, что слова с такими-то окончаниями обладают определенным свойством, то обратный словарь, в котором содержатся списки одинаково оканчивающихся слов, позволяет установить все слова, которые обладают тем или иным свойством, а также те, которые этим свойством не обладают. На основе обратного словаря могут быть получены списки слов, относящиеся к одному словоизменительному типу. Можно также выявить все слова, имеющие одинаковое строение концов, но разные грамматические характеристики, и получить данные о соотношении между окончанием слова и его принадлежностью к определенному словоизменительному типу. Возникает возможность определять синонимию и омонимию формантов, их сочетаемость, количественные характеристики отдельных формантов и их системы.

Морфологический анализ в своем удельном весе в системе автоматической обработки текстов существенно зависит от типа анализируемого языка. Ясно, что флективные языки несут больше информации в морфологических формантах, чем языки аналитического типа, выражающие синтаксические отношения главным образом с помощью порядка слов.

Попытки классифицировать языки по их отношению к некоторому единому общему алгоритму морфологического анализа оказались неплодотворными, поскольку такой алгоритм обладал бы нулевой универсальностью. Система морфологических признаков (декларативные знания) тесно связана с системой знаний процедурных — самим алгоритмом.

## 3.2. Проблемы слова. Вычислительная лексикография

По аналогии с предыдущими разделами следовало бы написать: «машинная лексика» или «уровень лексического анализа». Однако поскольку современный машинный лексический анализ начинается и кончается машинными словарями, имеет смысл уже на этом этапе рассмотрения выделить машинный анализ лексики в специальную лингвистическую науку — вычислительную лексикографию. Дополнительным резоном сделать это служит тот факт, что лексика является, так сказать,

первичным, основным слоем языка, составляющим сферу коммуникации, подверженным тенденциям развития, в первую очередь испытывающим на себе изменения в сфере коммуникации, реагирующим наиболее четким, формальным и быстрым образом на воздействия общества и технологии на язык.

### 3.2.1. Традиционная и машинная лексикография

Современная языковая ситуация, коммуникация с ЭВМ, потребности записи человеческого знания в память компьютера прежде всего и в огромной степени изменили лексический мир языка. Поэтому можно считать, что здесь уже сложилась и развивается целая лингвистическая наука.

Следует сразу разграничить два аспекта проблемы, хотя они в целом безусловно связаны. Компьютерная техника широко проникает в традиционную сферу лексикографии, в составление словарей, заменяя собой традиционную ручную картотеку и привнося новые методы и возможности в веками установившуюся технологию. Однако по целям применения в такой ипостаси она сводится к созданию словарей, предназначенных для человека. В то же время целый ряд разработок связан с использованием в интеллектуальных системах специальных машинных словарей естественных языков. В этих системах роль словаря огромна — от распознавания лексических единиц на уровне ввода и морфологического анализа и до моделирования элементов понимания и мышления. Таким образом, мы разграничиваем, в методологических целях, вычислительную лексикографию как часть обычной лексикографии, занимающейся составлением обычных, бумажных словарей для пользования человеком, и вычислительную лексикографию как науку о машинных словарях естественных языков, используемых в памяти компьютера для решения задач, требующих искусственного интеллекта.

Возможности ЭВМ в традиционной лексикографии чрезвычайно велики, поскольку они избавляют лексикографа от огромного многолетнего труда по сбору, расписыванию на карточки и анализу лексического материала. Отметим лишь одну из работ, воплощающих возможности машинных фондов

естественных языков для лексикографии. Можно создать справочник, в котором все словники наиболее значительных словарей были бы объединены. Таким справочником является «Сводный словник словарей русского языка» (Рогожникова, 1986). С помощью такого словника можно решать большое количество разного рода теоретических и практических задач. Так, можно выявить наиболее употребительную лексику современного русского языка и отобрать ее в словари различных объемов. Это важно при составлении толковых словарей, общего и учебного типа, для отбора лексики в русскоязычные словари других типов и пр. Поскольку в «Сводном словнике» отражены варианты слов, отмеченные в словарях, а также омонимы и некоторые грамматические особенности слов, он может оказать существенную помощь в научных исследованиях и учебных занятиях по лексикологии, служить базой для отбора лексики по различным семантическим, грамматическим и другим группам, для изучения омонимов, вариантов слов и т. п. По типу сводного словника словарей могут быть созданы словники словарей предшествующих эпох, терминологических словарей и пр.

Применение интерактивных методов позволяет исследователю-лингвисту использовать ЭВМ даже в тех случаях, когда какие-то языковые факты трудно или невозможно формализовать: при помощи диалога с компьютером появляется возможность найти достаточно обоснованные и приемлемые приблизительные решения.

В разделе настоящей работы о машинных фондах будет дополнительно рассмотрен вопрос о возможностях компьютерной лексикографии применительно к традиционному человеческому труду по составлению словарей. В данном же разделе мы сосредоточим внимание прежде всего на вопросах лексикографии машинной, описывающей особенности словарей естественного языка «внутри» компьютера.

В работе «Вычислительная лексикография» (Марчук, 1976) были определены основные отличия машинных словарей от обычных. Кратко говоря, эти отличия заключаются в следующем: если обычный, традиционный, «бумажный» словарь комплементарен к знаниям, имеющимся у человека-пользова-

теля, т. е. он дает некоторую дополнительную информацию к той, которая у человека уже есть, то машинный словарь в функции информирования автономен, т. е. он должен содержать всю необходимую для «искусственного интеллекта» машины информацию. Машинный словарь в применении в автоматизированных системах не может рассчитывать на то, что компьютер располагает «фоновым» знанием, необходимым для пользования словарем. Один из примеров того, как фоновое знание меняется от эпохи к эпохе, дает русско-французский словарь Н. П. Макарова. Русское слово «грошевик», означающее монету в две копейки, отсутствует даже в «Советском энциклопедическом словаре» (Макаров, 1908), (Советский энциклопедический словарь, 1980).

Машинный словарь должен содержать всю информацию, необходимую для работы с данным словом. Всякая неопределенность, неоднозначность и пр. сохранятся в выдаче, если в алгоритме работы с текстом не будут предусмотрены соответствующие процедуры.

Электронный словарь может объединять мультимедийные технологии и способы распознавания устной речи. Дж. Коккинакис в своей обзорной статье суммирует современные возможности словарной работы следующим образом. В течение веков лексикография была весьма трудоемким, требующим много времени и дорогим занятием, связанным с поисками разного рода примеров употребления лексем, исследованиями словарей и текстов, при этом охват всех аспектов языка был практически невозможен. Современная компьютерная технология коренным образом изменила характер работы лексикографов. Возможность разносторонней обработки больших объемов текстов за короткое время позволила получить огромное количество информации об употреблении слов, словоформ и словосочетаний, об особенностях грамматики, синтаксиса и пр. Кроме того, появилась возможность проследить эволюцию языковых элементов. В результате теперь можно получить настоящие современные словари за короткое время и сравнительно дешево.

Второй важный аспект касается формы словарей. При использовании CD-ROM с памятью в миллиарды байтов и при бы-

стродействующих компьютерах открылась возможность использовать мультимедийные технологии, и электронные словари снабжаются голосом, графикой и изображениями, что позволяет более эффективно воспринимать информацию. Наконец, весьма важным событием, повлиявшим на распространение и восприятие информации, явился Интернет. Словари, размещенные в базах знаний в разных концах мира, оказываются доступными самым различным пользователям, где бы те ни находились.

В настоящее время существуют три типа словарей:

- обычные бумажные словари;
- электронные словари с текстом на CD-ROM и с размещением в Интернете, т. е. электронные версии обычных бумажных словарей;
- мультимедийные словари, объединяющие текст, речь и изображение.

Следующим шагом будет объединение речевых и лингвистических технологий в мультимедийных словарях, что позволит резко увеличить их возможности. Это следующие технологии:

- лемматизаторы, дающие возможность распознать лемму в любой из форм. Это особенно важно для языков синтетического строя, богатых формами словообразования и словоизменения;
- морфологические анализаторы, которые дают возможность представлять лемму во всех ее словоизменительных разновидностях;
- конвертеры «графема — фонема», которые дают возможность усвоить произношение любого написанного слова;
- конвертеры «фонема — графема», которые приводят написание любой устной формы слова;
- корректоры написания слов (спеллеры), которые позволяют исправлять неправильное написание слова;
- синтезаторы «текст — речь», которые могут обратить любой текст в устную речь. В результате пользователь может услышать произношению любого слова в речи, помимо звучания этого слова в отдельности;
- устройства, распознающие устную речь и преобразующие ее в текст.

Все это безусловно расширяет возможности использования словарей для человека (Kokkinakis, 2001).

### 3.2.2. Отличия машинного словаря от обычного

Это основное функциональное отличие машинного словаря от предназначенного для человека; оно определяет состав и структуру первого. Машинный словарь, который работает в системе автоматической обработки текстов, имеет принципиальную переменность состава и может быть включен в работу практически с нулевым количеством слов, что бессмысленно для «человеческого» словаря. Структура словарной статьи обычного и машинного словаря также различна, она определяется целями и задачами машинной обработки текстов, в рамках которой она действует.

Следует сразу сказать, что машинный словарь в автоматизированной системе работает в условиях некоторой неопределенности, поскольку правильная идентификация нужной информации, привязанной к слову, достигается далеко не всегда. В этом обстоятельстве есть некоторая аналогия с процессом мышления человека. В умственной деятельности при поиске, например по аналогии, человек последовательно приближается к искомому представлению. «Поиск аналогии можно представить себе состоящим из следующих этапов: 1) перевод описания опытных данных с языка опыта на язык моделей; 2) выбор модели, наиболее близкой по описанию к описанию опытных данных; 3) "приведение в действие" модели и проверка соответствия выводов модели всему набору рассматриваемых новых экспериментальных данных» (Веденов, 1988, с. 104).

Перевод с одного языка на другой, по мысли физика, представляет собой поиск близкого по значению слова: стимулом здесь является совокупность слов, описывающих на языке опыта наиболее характерные, важные черты набора экспериментальных данных, а системой образов — совокупности соответствующих слов на языке модели. Уже на первом этапе может проявиться известная из практики неоднозначность аналогии: перевод с одного языка на другой неоднозначен потому, что данному описанию фактов на языке опыта может со-

ответствовать несколько различных описаний на языке моделей. Если мы разделим все известные нам модели на несколько классов (например, относящиеся к различным областям знания), каждый из которых имеет свой язык, то такая многозначность перевода становится неизбежной.

Неоднозначность, возникающую при использовании каждого лингвистического образа (слова) для отождествления связанных с этим словом понятий (одного или нескольких), можно расценивать как ошибку распознавания и последовательно применять методики для устранения ошибок идентификации. Этот вопрос можно трактовать в рамках общих проблем коммуникации «человек — машина». «Основной принцип коммуникации "человек — машина" на естественном языке... состоит в переводе конструкций, куда входят элементы естественного языка, в понятные машине коды. Для такого перевода в памяти машины (как часть машинной информации) содержится специальная лингвистическая информация, элементы которой представляются в виде словарей, списков наименований понятий, наборов фраз и т. п. Машина по лингвистической информации осуществляет перевод конструкций, куда входят языковые элементы коммуникации и соответствующие машинные коды... В общем случае механизм перевода состоит в том, что на каждом этапе поступающие элементы коммуникации сравниваются с набором альтернативных версий о том, каким должен быть поступающий на этом этапе элемент после поступивших ранее элементов. Результатом работы этапа считается наиболее подходящая версия из набора версий этого этапа...

Таким образом, на промежуточных этапах машина готовится воспринять не любой элемент естественного языка, а только те, которые соответствуют постепенно уточняющейся теме: поиск осуществляется не среди всех элементов лингвистической информации машины, а лишь среди тех, которые представляют адекватную альтернативу для данного этапа коммуникации» (Котов и др., 1988, с. 18–23).

При таком многоступенчатом машинном устранении неточностей особую значимость приобретает количественная мера в оценке версий. Количественная мера сходства символьных последовательностей (как на уровне графем, так и на

уровне слов, составляющих предложение) может служить одним из вариантов дальнейшего развития предиката в языках программирования, когда в практических целях необходимо результат сравнения таких последовательностей сводить не к множеству логических значений «истина» и «ложь», а дать более дифференцированную оценку такого сравнения.

Современные технологии активно используются в задачах определения лексических значений для прикладных проблем поиска и классификации (Поляков, 2004). Значение лексическое определяется как та часть семантического значения состава слова, которая в противоположность грамматическому значению целых классов и категорий слов присуща лишь данной лексической единице. Лексическое содержание большинства полнозначных лексем неоднородно и представляет собой смысловую структуру, иерархическую соподчиненность отдельных значений или лексико-семантических вариантов слова. Это свойство организации лексической семантики называется полисемией, или семантическим варьированием слова. В зависимости от характера предметной или понятийной отнесенности слова значения могут быть прямыми и переносными, в зависимости от степени контекстуальной обусловленности — свободными, фразеологически связанными, конструктивно обусловленными.

Значения слов традиционно фиксируются в толковых словарях. В этом направлении за два прошедших века (XIX–XX вв.) в России и в СССР была проделана колоссальная работа, которая позволила сохранить и зафиксировать как диахронную, так и синхронную семантику русской лексики. Можно отметить прочные лексикографические традиции также и в других странах (США, Англия, Германия). В качестве альтернативы бумажным словарям в проекте WordNet были предложены и апробированы компьютерные методы фиксации значений, основанные на экспликации семантических связей между словами средствами информационных технологий (базы данных и базы знаний). Двуязычные и многоязычные словари, как бумажные, так и компьютерные, содержат сопоставительное описание лексических значений. В качестве примера плодотворного сотрудничества в области традиционной и компью-

терной отечественной лексикографии можно привести словарь «МультиЛекс» (Шемякина, 2002).

Остановимся кратко на описании этого словаря. Наиболее важной составляющей любого компьютерного словаря является лексическая база, на которой он построен. В основу «МультиЛекса» положены словари, которые считаются наиболее полными и авторитетными из имеющихся словарей соответствующей тематики. В отличие от предыдущих версий, которые подразумевали работу со словарем исключительно с компакт-диска, «МультиЛекс 3.5.» можно полностью установить на жесткий диск, что заметно сокращает время поиска по словарю. Компания «МедиаЛингва» выпускает также ряд специализированных словарей, которые по желанию пользователя могут подгружаться в имеющуюся программную среду «МультиЛекса 3.5» (эти словари должны быть версии не ниже 3.5). На рынок уже выпущены:

- МультиЛекс 3.5. Английский. Экономика и право, в состав которого входят электронные версии англо-русского словаря по экономике и финансам под ред. А. В. Аникина (содержит 75 000 слов и выражений); англо-русский юридический словарь С. Н. Андрианова и др (содержит около 50 000 терминов); русско-английский юридический словарь И. И. Борисенко и др. (содержит 22 000 терминов);
- МультиЛекс 3.5. Англо-русский банковский и экономический словарь Б. Г. Федорова (содержит более 15 тыс. терминов).

Таким образом, пользователи «МультиЛекса 3.5.» имеют возможность самостоятельно создавать библиотеку словарей, которые наиболее подходят для их профессиональной деятельности. Все эти дополнительные словари могут быть установлены как на жестком диске, так и работать с помощью CD-ROM. Важно отметить, что несмотря на электронную форму словарей, пользователю доступна вся информация, имеющаяся в оригинальных печатных словарях, в том числе и структурная. Кроме того, именно электронная форма словаря позволяет существенно ускорить поиск словарной статьи, а также открывает перед человеком принципиально новые возможности в работе над переводом текстов.

«МультиЛекс» значительно облегчает поиск перевода устойчивых словосочетаний и выражений. Мы привыкли к тому, что при работе с традиционными бумажными словарями необходимо, во-первых, правильно определить опорное слово в словосочетании или идиоме, а во-вторых, просмотреть всю, зачастую весьма объемную словарную статью этого ключевого слова, чтобы найти нужное нам выражение. Для пользователей «МультиЛекса» этой проблемы не существует, так как словосочетание или выражение является такой же единицей поиска, как и отдельное слово. В ответ на такой запрос словарь выдает список примеров, т. е. словосочетаний или предложений, в которых встретилась данная единица поиска. При этом заданный нами порядок слов в искомом выражении может не соблюдаться. Так, можно задать поиск русского идиоматического выражения «когда рак на горе свистнет». В ответ мы получим следующий список соответствующих английских словосочетаний: *once in a blue moon — очень редко, после дождичка в четверг, когда рак на горе свистнет (см. once II); when pigs begin to fly — после дождичка в четверг, когда рак на горе свистнет (см. pig II); tomorrow come never — после дождичка в четверг, когда рак (на горе) свистнет (см. tomorrow I).* В ответ на запрос словарь выдает список словарных статей, в которых встретилось искомое выражение. Выбранные статьи имеют непосредственное отношение к запросу, т. е. содержат его в своем заголовке или в зоне русских эквивалентов, (где могут даваться примечания на английском языке). Если же искомое выражение встретилось только в зоне примеров какой-нибудь словарной статьи, то заголовок последней не выдается в качестве результата поиска. Для таких случаев программой предусмотрен специальный блок под названием «Примеры», куда помещаются подобные выражения.

«МультиЛекс» сохраняет возможность осуществлять поиск по алфавиту. Для этого существует режим «Алфавитный поиск», при котором пользователю предлагается алфавитный список всех заголовков словарных статей, по которому можно передвигаться и выбирать нужные слова «вручную» (с помощью мыши или клавиатуры).

Одной из важнейших особенностей компьютерного словаря «МультиЛекс» является его обратимость, т. е. возможность перевода как с английского на русский, так и с русского на английский. Это достигается за счет того, что заданный поиск производится не только среди заголовков словарных статей, но и непосредственно в их текстах. При поиске русского слова по англо-русским словарям «МультиЛекс» выдаст список словарных статей английских слов, в которых встречается искомое русское слово (в качестве переводного эквивалента или его части). При поиске английского слова или выражения по русско-английским словарям ситуация будет обратной. В обоих случаях искомая единица будет выделена «маркером» в текстах словарных статей, что позволит с легкостью ориентироваться в найденном материале и выбирать нужные переводные эквиваленты.

Если на компьютере пользователя установлена звуковая плата, то при работе с «МультиЛексом» можно воспользоваться программой синтеза, которая позволяет услышать произношение заголовков словарных статей на английском языке, а также английских слов в текстах словарных статей.

При работе с «МультиЛексом» некоторая информация может оказаться для пользователя излишней. Эту возможность электронный словарь предусматривает и позволяет в определенной степени участвовать в выборе того, что представляется на экране. Если мы не нуждаемся в такой служебной информации, как примеры, ударение русских слов, транскрипция и пр., то на панели инструментов мы можем найти соответствующие кнопки и отключить показ ненужной информации, а в случае необходимости — восстановить ее показ.

Был проведен эксперимент по определению того, насколько «МультиЛекс» удобен и обеспечивает быстрый поиск информации для переводчика по сравнению с бумажными словарями. Оказалось, что при умении пользователя работать с компьютером пользование «МультиЛексом» примерно в два раза сокращает время работы переводчика. При слабом знании компьютерной технологии и отсутствии навыков работы на нем бумажные словари оказываются предпочтительнее. Принимая во внимание освоение компьютерных технологий основным

составом пользователей, можно безошибочно предположить перспективность использования компьютерных словарей типа «МультиЛекс».

Вернемся к основной проблеме прикладной лингвистики в части лексического уровня языка — разрешению лексической многозначности. Проблема разрешения многозначности является одной из самых сложных прикладных задач, связанных с лексическим значением в том его определении, которое приведено выше. Задача автоматического разрешения лексической многозначности была впервые сформулирована в рамках направления науки и технологии, связанного с созданием систем машинного перевода. В дальнейшем эта проблема стала одной из ключевых не только при создании систем машинного перевода, но и систем обработки естественно-языкового текста других назначений — информационный поиск, автоматическое реферирование, автоматическое индексирование текстов, экспертные системы и пр. Десятки научных коллективов и коммерческих организаций занимаются этой проблемой. На регулярной основе проводятся соревнования между действующими компьютерными программными системами, предназначенными для этих целей.

Разрешение лексической многозначности определяется прикладными целями систем, в рамках которых требуется такое разрешение. Наиболее высокие требования к разрешению лексической многозначности существуют в машинном переводе. Эксперименты с машинным переводом показали, что основная часть смысловой (семантической) информации заключается в лексике. Неправильный перевод слов, как отдельных, так и в словосочетаниях, решающим образом влияет на понятность текста и эквивалентность текста перевода. Наоборот, при правильном переводе слов и словосочетаний грамматические ошибки, например на уровне синтаксиса, в меньшей степени влияют на понятность текста. Поэтому именно в задаче машинного перевода требуется максимально полное разрешение многозначности лексем.

В области поисковых технологий проблема многозначности не носит критического характера, так как сравнительно низкое качество поиска часто сглаживается большими объе-

мами информации в сети Интернет. В задачах классификации текстов влияние многозначности проявляется в том, что при выборе в качестве базового признака классификации единичной лексемы многозначность ее понижает точность классификации текстов. Задача определяет методику решения проблемы. Описание современных технологических приемов разрешения лексической многозначности для системы тематической классификации текстов Rubryx дается в работе Полякова (2004).

### 3.2.3. Вычислительная лексикография

Из всего вышесказанного следует, что вычислительная лексикография, машинные словари в системе автоматической обработки текстов являются неотъемлемой частью логического процесса автоматического распознавания смысла. Обладая конструктивной автономией как часть системы, словарь в то же время своей словарной лингвистической информацией тесно взаимодействует с автоматической процедурой решения той или иной информационной задачи.

Наиболее интересным и сложным видом обработки текста является перевод. В рамках машинного перевода используются наиболее сложные словари в том смысле, что словарная информация может быть самого различного состава, в зависимости от идеи машинного перевода: она может как охватывать все возможные коннотации слова, его синтаксические и семантические характеристики в словосочетаниях и пр., так и быть весьма ограниченной, необходимо включать только самые основные элементы статической и динамической информации для участия в алгоритме машинного перевода.

Может быть несколько другое деление вычислительной лексикографии на машинную и немашинную. Л. И. Колодяжная делит лексикографию на традиционную и вычислительную. Предметом первой является теория и практика составления словарей, предметом вычислительной (машинной) являются автоматические словари — или получаемые с помощью ЭВМ, или функционирующие на ЭВМ как автоматизированные системы, включающие массивы словарных статей и ком-

плекс обслуживающих программ. В вычислительной лексикографии при этом можно выделить три основные направления: а) автоматическое получение из текста с помощью ЭВМ различных словников, частотных словарей, конкордансов и пр.; б) теоретические и практические аспекты составления машинных словарей, для систем искусственного интеллекта, машинного перевода и пр.; в) создание словарей, являющихся машинными версиями традиционных.

Подобно автоматическим словарям в системах искусственного интеллекта и машинного перевода, автоматический словарь в книжной форме является совокупностью словарной базы данных и комплекса обслуживающих программ. Исходным материалом для создания машинной словарной базы являются тексты книжных словарных статей, поэтому особую актуальность приобретает задача автоматического анализа словарной статьи (Колодяжная 1987).

Таким образом, говоря о проблеме слова в вычислительной лексикографии, можно отметить, что три основных подхода посвящаются этой теме: лексицентрический — в рамках которого слово (словоформа) занимает главное место; текстоцентрический — в котором слово, оставаясь главной единицей анализа, определяется через текст; и словарноцентрический — при котором основная информация о слове получается в результате анализа словарной статьи имеющихся словарей, в том числе и машинными методами.

### 3.2.4. Словарноцентрический подход

Начнем со словарноцентрического подхода, поскольку он представляется самым естественным: каким же образом можно изучать слова, как не с помощью словаря (словарей).

Цели и задачи, которым служили и служат словари в течение многих веков, чрезвычайно разнообразны. В человеческом обществе словари осуществляют:

- регистрацию в словесной форме объективных данных о внешнем мире в свойственной данному языку и эпохе форме восприятия мира (энциклопедические и толковые словари);

- упорядочение в понятийно-словесной форме субстанции содержания данного языка (идеологические словари, словари-тезаурусы);
- нормализацию словоупотребления с целью облегчения языковой коммуникации (нормативные и терминологические словари);
- систематизацию лексического материала для обучения языку (учебные словари);
- перевод с одного языка на другой или другие языки (двуязычные и многоязычные переводные словари);
- прочие вспомогательные операции для пользования языком (словари-справочники, номенклатуры, специальные словари и пр.).

Л. В. Щерба выдвигал следующие основные противоположения для создания общей типологии словарей: 1) словарь академического типа — словарь-справочник; 2) энциклопедический словарь — общий словарь; 3) тезаурус — обычный толковый или переводной словарь; 4) обычный толковый или переводной словарь — идеологический словарь; 5) толковый словарь — переводной словарь; 6) неисторический — исторический словарь. Типология словарей может строиться и на других основаниях, например на противопоставлении активного и пассивного аспектов в процессе кодирования и декодирования сообщений. Могут быть приняты и другие принципы для типологии словарей, в зависимости от решаемой теоретической или практической задачи.

В рамках словарноцентрического подхода в вычислительной лексикографии разработаны специальные таксономии для лексических данных. Так, в одном датском проекте, целью которого является систематизация имеющихся словарей и приведение их по возможности к единому лексикографическому формату, определены два подхода. В терминологических исследованиях, когда разработка обычно начинается с концепта и описывает этот концепт с помощью отражающих его слов и выражений, имеет место так называемый ономасиологический подход. В наиболее общих лексикографических исследованиях разработки ведутся от слов и выражений, значение и употреб-

ление которых описываются. Это — семасиологический подход. Соответственно этому таксономия для лексикографических данных имеет следующий вид:

- Информация, определяемая проектом:
  - язык (естественный язык, диалект, социодиалект);
  - подход (семасиологический, ономасиологический).
- Общие типы информации:
  - этимологический (оригинальное, заимствованное);
  - графический (орфография, графические символы);
  - грамматический (части речи, инфлекция, словообразование);
  - фонетический (просодика, сегментация);
  - прагматический (контекстная, использование, оценочная и пр.);
  - семантический (субъект, семантические отношения, валентность и пр.).

Описанная таким образом система приводит к широкому диапазону словарей. Найден удобный формат для академических словарей, составлен словарь терминологического характера для широких предметных областей. Словарь может также использоваться как регистратор научных достижений, сведений о состоянии науки и т. д. (Descriptive tools, 1987).

Для современного состояния вычислительной лексикографии характерно, таким образом, то, что сохраняя свои прежние классификационные особенности, сформулированные в 1976 году, она значительно расширила арсенал своих средств, использовала многие дополнительные возможности. Машинные словари систем автоматической обработки текстов на естественных языках имеют широкое применение, фундируют использование различных алгоритмов, преодолевают расхождение между статикой и динамикой, действуя в рамках противоречия между статическим описанием лексики и необходимостью динамического ее анализа в текстах, каждый раз несущих новый смысл. Еще теснее связи ее с лексикографией традиционной, которая также расширила свои возможности и лучше удовлетворяет требованиям составления «человечески» ориентированных словарей и другой книжной продукции, свя-

занной с систематизацией лексики как самого подвижного и наиболее остро реагирующего на изменения слоя языка

Ю. В. Рождественский подчеркивает особую роль словаря-тезауруса в современной коммуникации, информатизации общества, в образовании. «Тезаурус — основа информатизации образования. Всякий тезаурус имеет основную рубрикацию, являющую принцип представления его содержания» (Рождественский, 2003, с. 111). Представляет интерес современный проект объединения русских словарей (энциклопедий, справочников) и текстов. В качестве конструкт-системы (посредника) между текстами и словарями конструируется ГИЗАУРУС (гипертекстовый тезаурус). В данном случае мы имеем дело с объединением словарноцентрического, лексицентрического и текстоцентрического подходов к описанию лексического состава языка.

ГИЗАУРУС (открытая гиперсистема) разрабатывается в рамках исследовательского проекта «Гипертекстовый генеральный свод лексики русского языка», поддержанный Российским фондом фундаментальных исследований и федеральной целевой программой «Русский язык» (Лесников, 2002). Автор трактует его как очередной виток спирали развития Машинного фонда русского языка в Интернете. Это система использования новейших информационных технологий для анализа (переработки в широком смысле слова) лексикографической информации в нелинейной форме в интерактивном режиме на ЭВМ с учетом иерархических, ассоциативных, сетевых и реляционных парадигматических связей посредством синтагматически реализованных в компьютерной форме (фреймов, слотов, фасет, шпаций, доменов, объектов, структур, узлов, указателей, сегментов, агрегатов, векторов, записей, констант, переменных, идентификаторов, списков, множеств, кортежей, наборов, ссылок, файлов, групп, полей, массивов, таблиц, предикатов, отчетов, шаблонов, этикеток, карточек и т. п.) оцифрованных лексикографических данных: текстовых, словарных, графических, аудио и видео, анимационных и т. д.

На компакт-дисках предполагается разместить толковые словари русского языка В. И. Даля и С. И. Ожегова, словари русского языка: антонимов, грамматический, морфемный, обратный словарь, словарь омонимов, орфографический, палин-

дромов, синонимов, словообразовательный, частотный, этимологический, энциклопедические словари Ф. А. Брокгауза и И. А. Эфрона, лингвистический, литературный, советский, логический, многие другие словари и справочники, учебники, хрестоматии и энциклопедии русского языка в виде Web-сайта, а также художественные произведения русских классиков XIX–XX вв. (Жуковский, Пушкин, Лермонтов, Гоголь, Чернышевский, Достоевский, Тургенев, Салтыков-Щедрин, Толстой, Чехов, Блок, Гумилев, Есенин, Платонов, Булгаков, Набоков и другие), историков (Карамзин, Соловьев), поэтов (Ахматова, Бальмонт, Баратынский, Барков, Державин и др.).

Проект такого рода позволит на основе информационных технологий объединить лексикографические материалы и обеспечить их оперативный ввод в научный оборот с целью оптимизации фундаментальных научных исследований и современной отечественной филологии.

Лексицентрический и текстоцентрический подходы описываются ниже, в последующих разделах, вместе с теми операциями, которые необходимы для их осуществления.

## 3.3. Лемматизация. Машиночитаемые словари

В связи с тем, что в подавляющем большинстве языков существуют изменения в форме слов, обусловленные правилами грамматики, для работы автоматического словаря необходимы алгоритмы лемматизации, т. е. приведения разных текстовых форм слова к его канонической, либо зарегистрированной в словаре как таковая, либо — при включении всех словоформ в автоматический словарь словоформ — назначенной в качестве главной, исходной, несущей основной состав информации.

### 3.3.1. Лемматизация

Операции, образующие лемматизацию, в принципе мало чем отличаются от автоматического морфологического анализа. Разница состоит в конечном продукте и цели. Если операционный морфологический анализ автоматизированного типа

имеет целью дать о слове наибольшую возможную информацию, независимо от уровня языка, т. е. морфологическую, синтаксическую, семантическую, лексическую и пр., то лемматизация, при всем практически повторяющемся наборе средств для ее реализации, предназначена лишь для установления связи анализируемой словоформы с каким-то одним ее каноническим представлением, зафиксированным в исходном словаре со всей необходимой информацией.

Лемматизация и машиночитаемые словари как практическая проблема связаны с важными теоретическими аспектами лингвистического характера. В капитальном труде К. Хеса, Дж. Брусткерна и В. Лендерса «Машиночитаемые словари немецкого языка», в котором освещаются не только собственно немецкие словари, но дается также подробное описание многих проблем машинного словарного дела, приводятся следующие конкретные проблемы машинной лексикографии и машинного определения слова.

Первая проблема связана с определением лексической единицы: какая лингвистическая сущность может считаться лексической единицей.

Вторая проблема: какова речевая реальность, речевая среда, в которой проявляется лексическая единица.

Третья проблема: каким образом из условий проявления данной лексической единицы и из ее речевых связей установить типологию лексических единиц.

Четвертая проблема связана с установлением связей лексической единицы на всех уровнях языковой системы, а также в сфере собственной лексической микроструктуры.

### 3.3.2. Составление машинных словарей

Методика изложения аспектов основных проблем машиночитаемых словарей и идентификации лексических единиц (лемматизации) обычно укладывается в следующую систему: а) определение (дефиниция) лексической единицы; б) корпус текстов — исходный массив текстов, в котором определима данная лексическая единица; в) типология машинных словарей упорядочивающих лексические единицы; г) лексические связи.

Важно отметить, что все эти аспекты рассматриваются применительно к конкретным задачам обработки естественно-языковых текстов и текстов словарей. Поэтому выделяют обычно следующие основные прикладные информатические задачи, в рамках которых происходит рассмотрение названных выше аспектов и проблем: машинный перевод, системы информационного поиска, вопросно-ответные и диалоговые системы, терминологические банки данных (HeSS et al., 1983).

Система соответствующих алгоритмов приводит к тому, что на основе обследования представительных массивов текстов для каждой лексической единицы выявляется ее микроструктура, характеризуемая следующими данными: 1) представление: лемма, произношение, набор лексических парадигм, орфографические варианты, способы сокращений; 2) объяснение: грамматические данные, парадигматические отношения, синтагматические отношения, фразеологические связи, стилистические, ареальные и прочие особенности, словообразовательные характеристики; 3) демонстрация: основные особенности употребления, статистические характеристики, библиографические отсылки.

В данной работе Хеса, Брусткерна и др. рассматриваются наиболее общие характеристики и особенности лемматизации как операции, производимой над совокупностью словоформ в заданном корпусе текстов. Для каждой из специфических информатических задач лемматизация имеет свои варианты. Так, для машинного перевода лемматизация есть часть морфологического анализа и приведения разных текстовых словоформ к одной канонической. Для информационного поиска лемматизация связана с опознанием лексических единиц одного семантического содержания. С этой точки зрения она является частью процесса индексации и поиска в тезаурусах, если тезаурус предусмотрен в системе.

## 3.4. О роли и функциях словосочетания

### 3.4.1. Словосочетание

В традиции общего языкознания различают обычно устойчивые, повторяющиеся сочетания слов, и сочетания переменные, свободно создаваемые в процессе речи (Маслов, 1987).

Хорошо известно из практики, что значения слов чаще всего определяются сочетаниями данного слова с другими словами в текстах и словарях. Единицей анализа в большинстве случаев выступает словосочетание, в том смысле, что именно оно несет чаще всего некоторый достаточно четко выделяемый фрагмент смысла. Поэтому роль словосочетаний в автоматическом анализе и синтезе естественно-языкового текста достаточно велика. Так, в терминологических словарях до 50 % входных лексических единиц представлены словосочетаниями. Роль и функции словосочетаний в терминологических текстах будут рассмотрены ниже в разделе, посвященном терминологии и терминоведению.

Прежде чем рассматривать особенности словосочетаний в подлежащих компьютерной обработке информатических текстах, приведем некоторые чисто лингвистические соображения относительно языковой природы словосочетаний и их роли среди лингвистических единиц других уровней языковой структуры.

### 3.4.2. Классификация словосочетаний

Известный русский лингвист, внесший значительный вклад как в общее языкознание, так и в прикладное, Борис Николаевич Головин, так определяет словосочетание: это два или несколько полнозначных слов, объединенных *одной* синтаксической связью. Примеры: *видеть птиц, встречать товарища, идти вперед, петь весело, кричать и разговаривать* и т. п. Классификационные деления словосочетаний не вполне ясны. Классификация словосочетаний осложняется, как говорит Б. Н. Головин, вмешательством предложения в их структуру. Предложение преобразует словосочетания, например делает их предикативными. Внутри словосочетания возникает сложная борьба между морфологией, лексикой и синтаксисом. Так, морфология позволяет присоединить наречие к глаголу, но лексика не разрешает сказать, например, *читать вплавь*. Словосочетание — это явление, переходное от морфологии к синтаксису: словосочетание морфологично, поскольку в нем реализуется морфологическое свойство слова присоединять к себе

другие слова или присоединяться к ним; словосочетание синтаксично, поскольку в нем возникают необычные для морфологии связи и отношения, на основе которых формируются члены предложения, обособление, синтаксическая однородность и другие синтаксические явления (Головин, 1966). Рассуждения Б. Н. Головина важны в том смысле, что в них мы видим точку зрения человека, сочетающего традиционные лингвистические подходы к проблеме со специфическим взглядом эксперта в точных методах языкознания.

Рассматривая словосочетание с точки зрения порождения текста, другой видный советский лингвист, Александр Васильевич Зубов, отмечает, что на втором этапе порождения текста, на котором совершаются процессы отбора лексических единиц и грамматических форм, основной единицей выступает именно словосочетание. В отличие от слова оно выражает некоторые отношения между элементами реальной действительности. В отличие от предложения оно никогда не выступает относительно законченным сообщением и не соотносит содержащуюся в нем информацию с тем или иным объективно-модальным или временным планом. С обучением жизни и языку в памяти человека запоминаются также и наиболее употребительные словосочетания. Выработанная связь между определенными элементами ситуаций и словосочетаниями все более закрепляется, и практически все словосочетания всегда представляют собой социолингвистически обусловленную категорию. Любой автор в основном повторяет, воспроизводит те сочетания, которые были до него сделаны в данном обществе в соответствии с потребностями общественной жизни и которые он слышал в аналогичных контекстах. Иными словами, в качестве элементов структуры текста выступают не предложение и не слово, а предмет и его признаки, которые могут быть выражены различными языковыми средствами. Чаще всего для этой цели используются различного типа словосочетания. Набор этих словосочетаний индивидуален для каждого человека и зависит от многих факторов. В частности, для взрослого человека такой набор во многом определяется профессией, социальной принадлежностью, окружающей средой. Возможно, конечно, и появление в тексте совершенно новых словосочетаний, но это, как правило, прерога-

тива художников слова — писателей, поэтов, журналистов и т. п. (Зубов, 1996, с. 107).

Психологические исследования по восприятию речи показывают, что для читающего поток письменной речи расчленяется на отдельные связанные по смыслу комплексы, которые при чтении воспринимаются на уровне словосочетаний. Эта мысль подтверждается лингвистическим и методическим материалом. К этому же выводу приводит и ритмико-просодический анализ процесса порождения высказываний. В то время как на уровне предложения его минимальной единицей является слово, на уровне абзаца его минимальной единицей является словосочетание, выступающее как сложная номинативная единица, эквивалентная слову.

Эти идеи находят подтверждение и в исследованиях специалистов по квантитативной лингвистике. Элементами словаря языковой системы, по их мнению, могут считаться такие единицы, которые могут храниться и воспроизводиться большинством носителей языка в целостном, готовом виде. Сюда войдут все не очень длинные слова, много словосочетаний-клише и некоторое число коротких предложений.

Особое место занимают словари фиксированных словосочетаний, фразеологических оборотов или неразложимых единств. В свободных словосочетаниях смыслы членов (отдельных слов) как бы складываются и общее значение словосочетания ясно вытекает из значений составляющих. В свободном словосочетании налицо соответствие между членимостью формы и членимостью содержания. Фразеологизм подобен отдельному полнозначному слову. Эта эквивалентность не формальная, а смысловая и функциональная: фразеологизм выполняет в нашей речи ту же функцию, что и полнозначное слово, т. е. выражает понятие, обозначает явление, участвует в построении высказывания. Фразеологизм входит в нашу речь как готовый элемент языка.

Классификация фразеологизмов может быть различной и в большинстве случаев зависит от целей исследования. Так, Б. Н. Головин выделяет идиомы, сращения, фразеологические единства и фразеологические сочетания. Новейшие исследования выделяют, например, коллокации — «устойчивые слово-

сочетания», специфика которых несвободная сочетаемость слов-компонентов (Борисова, 1996).

Особенностью машинной обработки словосочетаний является включение их в специальный словарь, функционирующий отдельно от входного словаря самостоятельных лексических единиц. Так, например, в машинном переводе создается специальный словарь оборотов, который работает на начальных стадиях в общем алгоритме анализа и исключает из дальнейшего анализа, если это необходимо, идиомы, фразеологизмы, другие несвободные словосочетания, степень связанности элементов которых определяется их характеристиками в исходном конкордансе. При этом им приписывается вся необходимая для дальнейшего анализа информация, например информация об их синтаксической функции, лексическом значении и пр. (Марчук, 1983).

## 3.5. Автоматический контекстологический словарь

### 3.5.1. Теория детерминант

Разрешение лексической многозначности в прикладном плане является одной из наиболее актуальных проблем компьютерной лингвистики. Выше, в разделе 3.2.2., говорилось о значении лексической многозначности и ее роли в прикладных системах. Для машинного перевода как для наиболее важной проблемы вычислительной и компьютерной лингвистики разрешение лексической многозначности (неоднозначности) имеет большое значение, поскольку именно в переводе необходимо выдать точный эквивалент слова, необходимый в данном лексическом контексте.

Сразу заметим, что лексическая многозначность в прикладном плане не отделяется от лексической синонимии. Так, слово *коса* будет иметь, по крайней мере, три разных значения — «элемент прически», «изгиб реки» и «орудие труда», хотя в традиционных словарях эти значения могут быть поданы как лексические синонимы. Словоформы *червяк* как живой организм и как элемент механической передачи — термин — также могут

быть разведены по разным статьям словаря. Отметим, что и в рамках традиционного языкознания вопрос о лексических омонимах и их подаче в словарях достаточно сложен и допускает разные трактовки и решения. В прикладном же плане, с точки зрения автоматического разграничения многозначных слов и омонимов, граница между ними стирается подобно тому, как стирается разница между словоизменением и словообразованием в автоматическом морфологическом анализе.

Одной из актуальных реализаций идеи контекстной связанности слова является автоматический контекстологический словарь, с помощью которого в автоматизированных системах разрешается лексическая многозначность слова. Раздел об автоматическом контекстологическом словаре включен в часть книги, посвященную словосочетаниям, потому что в подавляющем большинстве случаев лексическая многозначность слова разрешается в пределах словосочетания.

Термин «автоматический контекстологический словарь» введен автором в 1976 году в публикации версии такого словаря для машинного перевода многозначных слов с английского языка на русский (Марчук, 1973; Марчук, 1976; Marchuk, 1979). В рамках разработки системы англо-русского машинного перевода АМПАР, которая началась примерно в 1960 году в научно-исследовательском институте Министерства радиопромышленности СССР и затем в 1974 году была передана для дальнейшей разработки и совершенствования во Всесоюзный центр переводов научно-технической литературы и документации Государственного Комитета СССР по науке и технике и Академии наук СССР, в течение нескольких лет группой исследователей и разработчиков составлялся такой словарь для машинного перевода газетных и публицистических текстов. Исходным материалом для составления такого словаря служили словари-конкордансы текстов на английском языке, для которых также существовали тексты переводов на русский язык, т. е. можно сказать, что работа велась с использованием параллельных текстов и по методике, которая в настоящее время «озвучена» в рамках теории и практики так называемой корпусной лингвистики. Исследовательская группа оказалась пионером в таком решении вопроса. Академическая наука того

времени не одобряла подхода, который ей представлялся кропотливым и трудоемким выискиванием «контекстных зацепок» в разрешении многозначности и более красивым теоретически выглядел подход с использованием универсальных смысловых множителей и априорным определением значений слов. Тем не менее практические потребности требовали не приблизительного, а точного определения лексических значений, и от «контекстных зацепок» невозможно было уйти без потери точности перевода. Другое дело, что алгоритмы запроса контекста нуждались в обобщении, и такое обобщение было сделано в теории детерминант, которая в 1976 году и была опубликована в работе автора данной книги (Марчук, 1976). Следует особо отметить роль в этой разработке руководителя научно-исследовательского и опытно-конструкторского коллектива кандидата филологических наук Юрия Александровича Моторина, который смело взялся за трудное дело исследования и разработки сложнейшей научной и прикладной проблемы и возглавил коллектив разработчиков (В. И. Щербинин, А. В. Княгинин, Е. Е. Ловцкий и др.). Существенный вклад в работу внес известный писатель Д. А. Жуков, который также в течение нескольких лет активно работал над проблемами машинного перевода. Теперь, когда проблема машинного перевода насчитывает уже почти пять десятков лет, можно уверенно оценить все то, что было тогда сделано. Как бывает в истории науки, многие концепции оказались тупиковыми и не дали положительных результатов в решении проблемы, другие даже уводили куда-то в сторону от нее и служили трамплином для погружения в море совершенно других идей и концепций. В то же время идея и результаты контекстного разрешения лексической многозначности и принцип контекстологического словаря показали свою действенность и дали возможность ввести системы машинного перевода с английского и немецкого языков на русский АМПАР и НЕРПА в промышленную эксплуатацию, что и было отмечено серебряной медалью Выставки достижений народного хозяйства в 1980 году. Надо также отметить, что реалистический подход к разрешению лексической многозначности, осуществленный в работах Всесоюзного центра переводов, получил полную поддержку со стороны Обще-

союзной группы «Статистика речи», руководимой выдающимся русским ученым в области прикладной и теоретической лингвистики профессором, академиком Международной академии информатизации Раймондом Генриховичем Пиотровским, многолетнее сотрудничество с которым в огромной степени обогатило автора настоящей книги. Роль Р. Г. Пиотровского в развитии и становлении структурной, прикладной, квантитативной, компьютерной лингвистики в СССР и в России отражена автором в работе, посвященной юбилею Р. Г Пиотровского (Marchuk, 2003).

Важным моментом является также и то, что концепция машинного перевода в модели переводных соответствий, сформулированная автором в докторской диссертации, защищенной в МГУ им. М. В. Ломоносова в 1980 году, выдержала проверку временем и машинной эксплуатацией и служит основанием практически для всех ныне действующих коммерческих и промышленно используемых систем машинного перевода, в той или иной ее разновидности, при сохранении принципов структуры, теории и композиции. Это подтверждает тщательное исследование П. Н. Хроменковым (Хроменков, 2000) действующих систем машинного перевода.

Теоретической основой контекстологического словаря является теория детерминант (Марчук 1973, Марчук 1979). Согласно этой теории, каждое значение (перевод) многозначного слова, отличающееся от наиболее общего значения (общего выхода алгоритма перевода), детерминируется в контексте другими словами, сочетаниями слов или другими текстовыми признаками, такими, например, как грамматические категории в их эксплицитном контекстном выражении. Эти определяющие перевод слова, группы слов, грамматические категории называются детерминантами.

Два основных теоретических положения лежат в основе практической работы по созданию контекстологического словаря. Первое заключается в том, что значение многозначного слова (лексическое его значение) приравнивается к переводу. Это означает, что слова-синонимы будут определяться как разные значения: так, «лингвистика» и «языкознание» будут разными значениями слова *linguistics*, и это английское слово будет рас-

сматриваться как многозначное. Второе положение заключается в том, что детерминанты и переводы (значения) слов выбираются из текстов. В этом и заключается текстоцентрический аспект подхода к созданию контекстологического словаря. Исходным материалом для составления словаря является словарь-конкорданс, составленный на определенном массиве текстов входного языка, для перевода которых создается система.

Словарь-конкорданс представляет собой список текстовых употреблений каждого слова, взятых в контексте определенного размера. Приведем пример конкорданса для английского глагола *to safeguard:*

*It is important* ***to safeguard*** *the quantities of rice proposed in ...*

*New tariffs on household goods* ***to safeguard*** *German customers will eventually...*

*... to establlsh organizations to help* ***safeguard*** *its work.*

*They proposed a most comprehensive and* ***safeguarded*** *disarmament plan ...*

*He will be personally* ***safeguarding*** *all above mentioned items*

и т. д.

В конкордансе все слова упорядочены по алфавиту, а внутри группы словоформ конкретной лексемы они также упорядочены по форме. На тот случай, если контекст не дает возможности правильно определить перевод слова, в каждой строке конкорданса указывается специальным образом закодированное место в массиве исходных текстов с тем, чтобы можно было обратиться к более широкому контексту.

При использовании параллельных текстов составляется также конкорданс по исходному массиву текстов переводов, и не составляет труда сопоставить каждое употребление словоформы во входном тексте с ее переводом в тексте выходном.

Такой словарь-конкорданс исчерпывающим образом иллюстрирует использование данной лексемы и все ее значения во входном языке, представляя ее контекстное окружение, что дает возможность выделить все детерминанты.

Можно себе представить, что в отношении какого-то конкретного слова конкорданс будет неполным, поскольку частотность смысловых слов в значительной степени зависит от стиля данного конкретного подъязыка. В этом случае, а также и в других сомнительных случаях, составитель алгоритма перевода многозначного слова по контексту имеет полную возможность обратиться к соответствующему словарю и ввести в алгоритм контекстные признаки других переводов слова, не зафиксированных в исходном конкордансе.

Степень связанности компонентов словосочетаний, отраженных в контекстологическом словаре, не играет особой роли, кроме тех случаев, когда неразложимые словосочетания или фиксированные обороты имеет смысл не включать в словарь, а рассматривать их отдельно в специальном словаре оборотов, который работает в системе до контекстологического словаря. Такое решение является целесообразным в связи с тем, что обороты со всеми их компонентами выполняют единую синтаксическую функцию и один фиксированный перевод, что избавляет от необходимости рассматривать и анализировать каждое составляющее оборот слово отдельно. Например, если анализировать каждую словоформу английского словосочетания *in case of* или *in accordance with*, то такие слова, как *in, case, of, with* дадут огромное множество переводов, когда мы рассмотрим конкорданс на каждое из них. Между тем в рамках фразеологического единства они выполняют строго фиксированную синтаксическую функцию и имеют один перевод — всего словосочетания.

Условиями работы контекстологического словаря, точнее, состав информации, которая необходима для его действия, являются:

- предварительная работа словаря оборотов — исключаются фразеологические единства;
- разрешение лексико-грамматической омонимии. Так, слово *round* может быть либо существительным, либо прилагательным, либо глаголом, а иногда наречием или предлогом. В контекстологический словарь для разрешения лексической многозначности (если таковая у него будет зафиксирована в конкордансе) это слово попадет как слово

соответствующей части речи (лексико-грамматического класса);

- наличие в информационной ячейке слова морфологических признаков. Контекстологический словарь работает после алгоритмов автоматического морфологического анализа и каждая словоформа имеет соответствующие морфологические признаки (числа, притяжательного падежа и т. п.);
- имеется система правил пропуска несущественных слов для поиска нужных детерминант.

Каждое значение многозначного слова в контекстологическом словаре определяется диагностирующими признаками контекста — детерминантами. Каждый алгоритм перевода имеет недетерминированный общий выход, который представляет собой результат работы алгоритма в том случае, если в контексте не нашлось детерминант, определяющих перевод в данном конкретном случае.

Однозначность или многозначность слова также определяется изучением исходного конкорданса. Однозначные (имеющие один перевод) слова не входят в контекстологический словарь и переводятся с входного на выходной язык простым списком, в котором каждому входному слову сопоставлен один переводной эквивалент.

### 3.5.2. Алгоритм перевода многозначного слова

Алгоритм перевода многозначного слова представляет собой последовательность запросов контекста на наличие в нем детерминант, каждая из которых одна или в некоторой совокупности с другими детерминантами однозначно определяет один перевод многозначного слова. Операции по анализу контекста и прочие действия, связанные с переводом, — приписывание слову перевода, обработка других слов при переводе — выполняются стандартными операторами, каждый из которых представляет собой некоторую подпрограмму для ЭВМ. Совокупность операторов, которых в системе АМПАР восемь, образует некоторый язык стандартных операторов, который используется для записи алгоритмов машинного перевода и мо-

жет быть применен и в других задачах автоматической обработки естественно-языковых текстов.

В контекстологическом словаре использованы следующие стандартные операторы:

1. Операторы поиска в тексте лексических детерминант (отдельных слов, списков слов или лексико-грамматических классов слов). Все операторы имеют следующую структуру:

**A_____B_____C_____D**

*Рис. 5.* **Структура стандартных операторов**

Часть **A** — код оператора.

Другие части структуры заполняются информацией в соответствии с назначением оператора.

Операторы поиска имеют следующие коды, каждый из которых обозначает определенное направление и характер поиска:

**rgt** — поиск вправо, причем **rg** — поиск вправо от детерминанты, найденной предыдущим оператором (поиск по цепочке);

**lft** — поиск влево, **lf** — поиск влево по цепочке.

Часть **B** — номер подпрограммы пропуска несущественных для данного поиска слов. При поиске детерминанты пропускаются несущественные для данного поиска слова, если они окажутся между переводимым словом и детерминантой. Например, при поиске последующего существительного необходимо пропускать прилагательные, являющиеся определением к искомому существительному, поскольку не они определяют перевод, а искомое существительное — детерминанта.

2. Оператор проверки слова на грамматический признак. Код оператора **ch**. Проверяться может или основное слово, выбранное алгоритмом, или детерминанта. В качестве признака может выступать морфологическая информация, полученная на предыдущем этапе морфологического анализа, который может быть совмещен с поиском по словарю,

например информация «множественное число» или «окончание s». В части **B** этого оператора содержится номер оператора, нашедшего слово, подлежащее проверке.

3. Оператор приписывания перевода. Код оператора **tr**. Он может приписывать перевод только основному слову, занося в его информационную ячейку номер перевода из словаря выходного языка или другого хранилища переводов, и может также переводить (в том числе давая и нулевой перевод) слова, входящие с основным в словосочетания разной степени устойчивости.
4. Оператор отхода к другим операторам схемы (алгоритма перевода). Это последний оператор из четырех, необходимых для перевода многозначных слов.

Все операторы нумеруются по порядку, и номер оператора является указателем на то, какое слово должно проверяться и к какому оператору нужно отходить при ответе «нет». При ответе «да» выполняется следующий по порядку оператор.

Алгоритм перевода многозначного слова в данной записи имеет следующий вид, показанный на рис. 6:

**BACK (глагол)**

| | | | | |
|---|---|---|---|---|
| 1 | Rgt | 3 | Down | 3 |
| 2 | Tr | | Отступиться | |
| 3 | Rgt | | Out | 6 |
| 4 | Rg | | Of | 6 |
| 5 | Tr | 3,4 | Уклониться от | |
| 6 | Rgt | 3 | Away | 8 |
| 7 | Tr | | Отказаться | |
| 8 | Rgt | | Up | 10 |
| 9 | Tr | 8 | Поддержать | |
| 10 | Rgt | 1 | (существительное) | 13 |
| 11 | Rg | | Up | 13 |
| 12 | Tr | 11 | Поддержать | |
| 13 | Tr | | Поддержать | |

*Рис. 6.* **Алгоритм перевода многозначного английского глагола *to back*, записанный в стандартных операторах**

Контекстологический словарь представляет собой, таким образом, полный набор алгоритмов перевода многозначных слов для всех частей речи. Он составлен по конкордансам для перевода слов в системе машинного перевода, хотя может работать и самостоятельно, при условии выполнения требований к предварительной информации к словам. Он также может использоваться в программированном обучении языку. Описанный контекстологический словарь для перевода многозначных английских слов на русский язык работает в системе машинного перевода АМПАР и системах серии СПРИНТ (Марчук, 1976, 1983).

Языковая пара может быть другая, принцип контекстологического словаря применим и к другим языкам. Контекстологический словарь подобного типа был составлен по конкордансам также и для немецко-русского машинного перевода (система НЕРПА Всесоюзного центра переводов научно-технической литературы и документации) (Марчук, 1983). Важно при этом отметить, что система программирования словарных статей контекстологического словаря может быть иной, в частности, она может отражать более совершенные методы программирования и использования естественно-языкового интерфейса, однако сам идейный принцип запроса контекста на наличие — отсутствие детерминант остается постоянным в любом варианте программирования, равно как и методика составления контекстологического словаря.

До настоящего времени нам неизвестны другие разработки, в которых был бы осуществлен в такой подробности принцип детального контекстного анализа. Необходимость разрешения лексической многозначности неоднократно подчеркивалась разработчиками и исследователями проблем компьютерной лингвистики. Существуют и другие способы ее разрешения (см., например, Поляков, 2004). Некоторые из них сводятся к ограничению предметной области, в рамках которой многозначность слов, в частности, слов-терминов, резко ограничивается или вообще устраняется. В таких задачах, как информационный поиск, требования к точности разрешения лексической многозначности не такие строгие, как в проблеме машинного перевода, и можно также ограничить лексическую

многозначность или удовлетвориться ее приблизительным решением. В некоторых системах машинного перевода значения многозначного слова даются некоторым перечнем в скобках в переводимом тексте для того, чтобы постредактор вычеркнул ненужные значения.

Все это делается потому, что исследование конкордансов, хотя собственно составление конкордансов выполняется на ЭВМ, требует участия человека-переводчика, как и создание алгоритмов перевода. Затраты труда на составление контекстологического словаря для представительной предметной области достаточно велики, требуются усилия коллектива в несколько десятков человек в течение нескольких лет, при максимальном использовании вычислительной техники. Поэтому многие современные фирмы, занимающиеся машинным переводом, решают эту проблему альтернативными способами, в частности и особенно ограничением предметной области. Создаются словари, рассчитанные на использование в переводах материалов, например, в областях финансов, робототехники, а также химических текстов и пр. Однако, как показывают исследования теоретиков терминоведения (см., например, Авербух, 2004) и практиков перевода научно-технической литературы (см., например, Борисова, 2002), предметные области часто пересекаются, и, соответственно, значения даже терминологической лексики, не говоря уже об общеупотребительной или общенаучной, можно определить только по контексту.

Можно также добавить, что пересечение значений слов, в частности слов-терминов, определяется такой трудно формализуемой характеристикой подъязыка, как его широта, политематичность, широта предметного поля. Точные исследования в этой области пока еще не получили широкого распространения, однако можно утверждать, что наиболее эффективное использование принципа контекстологического определения значений многозначного слова, в том числе и особенно слова-термина, который играет все более важную роль в современном языковом общении, будет связано с успехами в количественной — квантитативной или другой объективной и формальной методике определения характеристик широты предметного поля.

### 3.5.3. Контекстологический словарь как организатор базы знаний

Идея контекстологического словаря получила дальнейшее развитие как применительно к словарному делу, так и в более широком плане. Был составлен ряд словарей с использованием принципа контекстного анализа (Меркулова, 2000; Галяшина, 2003; Mehrak, 2002 и др.). А. Л. Семенов (Семенов, 1994) разработал направление, связанное с использованием контекстологического словаря для систематизации терминологии в многоязычных политематических базах знаний.

Любая современная база знаний политематична по своему определению. Напомним, что база знаний отличается от базы данных тем, что первая содержит не просто данные, а структурированные, определенным образом систематизированные и упорядоченные данные. С одной стороны, чем новее база знаний, тем сильнее в ней интерференция предметных (особенно смежных) областей, а с другой стороны — ценность базы знаний в ее постоянной новизне. Новизна базы знаний — одно из самых критичных условий ее существования. Политематичность базы знаний повышается с расширением ее масштабов.

Естественный язык, являющийся средой представления знаний в базе знаний, теряет специализацию, определяемую тематической направленностью предметной области, по мере расширения базы знаний. Понятие подъязыка предметной области, которое помогает решать целый ряд проблем прикладной лингвистики, в том числе и машинного перевода, как это показал В. М. Зелко на примере китайского языка (Зелко, 1991), по мере расширения базы знаний теряет свои разрешающие способности: лексический состав универсализируется, а терминосистема расширяется. При этом возрастает важность системности терминологии, обуславливающей политематическую базу знаний.

Процесс понимания, обобщения и генерации, и прежде всего аналитико-синтетический процесс возможен только на естественном языке, поскольку источником знаний в начале аналитико-синтетического процесса и адресатом представления знаний является человек. Разного рода искусственные среды, такие, как языки программирования, математические фор-

мулы и символы, язык химических реакций и т. п., безусловно важны в организации баз знаний и способствуют коммуникации на разных языках, но они не могут полностью заменить естественный язык как средство организации политематичной и многоязычной базы знаний. Естественный язык, таким образом, опираясь на свои лексические средства, в первую очередь на уровне терминологии, является универсальной средой представления знаний, в которую возможно или даже необходимо включение искусственных сред.

При представлении знаний в информационных системах требуется значительная формализация, ясность, точность и однозначность представляемой информации, т. е. требуется высокая формализация плана выражения. Это условие выполняется прежде всего посредством повышения терминологизации представляемой информации.

Суть функционирования базы знаний на естественном языке сводится к целенаправленной обработке информации, цель такой обработки имеет прагматический характер и сводится к компрессии информации. Пределом компрессии является отображение содержания информации ключевыми словами.

Под ключевым словом понимается слово или устойчивое словосочетание, выбираемое из некоторого текста, отражающее смысловое содержание явления или понятия, составляющего предмет описания. По законам формальной логики, чаще всего этим словом или словосочетанием должен быть термин предметной области, к которой относится текст.

Метод координатного индексирования, составляющий в настоящее время основу технологии информационной обработки текста, базируется на том, что основное смысловое содержание текста может быть с достаточной точностью выражено списком ключевых слов, которые явно или в скрытом виде содержатся в индексируемом тексте. В аналитико-синтетическом процессе в базе знаний взаимодействуют три вида текстов на естественном языке: текст-источник знания, текст информационного запроса и текст представления знания. По технологии текст-источник знания подлежит компрессии, в результате которой повышается значение использования терминологии предметной области текста. Текст информационно-

го запроса максимально приближен к списку ключевых слов. А текст представления знания в значительной степени повторяет тексты-источники знания. Таким образом, основной особенностью обработки информации на естественном языке при накоплении, хранении представлении знаний является ее компрессия, следствием которой является повышение терминологической насыщенности текста.

Для всесторонней лингвистической оценки слова, в частности и особенно термина, необходимо иметь в памяти как можно больше примеров использования этого термина в текстах. Определение термина, даваемое в терминологическом словаре в виде дефиниции, является важным элементом определения его значения, но отнюдь не полным и не исчерпывающим. Веронис в кратком определении выражает принципиальную точку зрения на это обстоятельство: «Смотри не на значение, а на использование» — Don't look for the meaning but for the use (Veronis, 2000). Поэтому чрезвычайно важно контекстное использование термина, отражаемое в примерах его использования.

Существенная разница между толковым словарем и базой знаний заключается в объеме хранимой информации, относящейся к термину. Толкование в толковом словаре выполняет функцию определения. Оно прежде всего связано с понятием и разъясняет его смысл, используя для этого самый термин и словосочетания с ним. Информация, накапливаемая в базе знаний, хотя она и связана с толкованием, в большей мере относится к контекстам, в которых участвует данное понятие. Эти контексты способны в достаточной полноте отразить все разнообразие лингвистической информации, необходимой для интерпретации термина.

Используя преимущество базы знаний в части хранения и презентации информации о понятиях предметной области и вместе с тем уделяя внимание лингвистической характеристике терминов, их толкованиям и обобщенным примерам употребления, можно создать, говорит А. Л. Семенов, фундаментальный контекстологический словарь какой-либо предметной области. Основное назначение такого словаря — это исчерпывающая презентация семантики термина косвенным образом,

например способом представления в ответ на терминологический запрос большого количества обобщенных примеров (контекстов) употребления данного термина.

Контекстологический способ семантического определения термина был использован А. Л. Семеновым при разработке терминологии предметной области маркетинга. Это относительно новая в языковом отношении (с позиций русского языка) предметная область характеризуется высокой сложностью определения основных понятий и подбора эквивалентов для иноязычных терминов. При таком положении определение термина способом демонстрации его в различных контекстах во многих случаях является единственным способом представления его в словаре. Преимуществом контекстологического способа определения термина является, с одной стороны, наглядная иллюстрация функционирования термина в тексте, а с другой — отображение его семантических связей с другими терминами данной предметной области. Таким образом, в контекстологическом словаре наиболее продуктивно используется свойство системности терминологии.

В методическом плане преимуществом контекстологического словаря по сравнению с другими лексикографическими источниками, ориентированными на системный подход при отображении терминологии, является то, что, используя идеологию базы знаний в части хранения и презентации информации, контекстологический словарь снабжает участника аналитико-синтетического процесса (например, переводчика), моделями текстов, которые могут быть использованы как при понимании входного текста (оригинала), так и при генерации выходного текста (перевода). То есть в теоретическом плане контекстологический словарь в части идеологии своего организационного построения и фактического содержания полностью согласуется с идеологией процессов восприятия, хранения информации, мышления и разработки языковых форм общения, на которые ориентированы теоретические принципы построения баз знаний, поскольку чаще всего источником для этих теоретических принципов служит тот факт, что человек, познавая новую ситуацию, выбирает из своей памяти некоторый известный ему образ (например, фрейм) и затем, сохраняя объ-

единяющие эти ситуации сведения и изменяя различные детали, создает новый образ (фрейм), пригодный для понимания новой ситуации и усвоения конкретного нового знания.

Как фрейм в структуре базы знаний служит для представления стереотипной ситуации, так и обобщенный пример в контекстологическом словаре может быть использован для представления одного и генерирования другого — стереотипного текста. В таком случае, если в контекстологическом словаре обобщенный пример приводится параллельно на двух или более языках, он становится особенно полезной, способной к замене и подстановкам моделью для такого аналитико-синтетического процесса, как перевод, поскольку вариант на одном языке может быть использован для понимания смысла (оригинала), а вариант на другом языке — для генерации текста (перевода).

Подобно фрейму, полная статья контекстологического словаря содержит информацию трех видов:

- собственно входное слово статьи (ключевой термин предметной области), которое необходимо для соединения с ситуацией, имеющейся в тексте и инициирующей поиск в контекстологическом словаре в процессе аналитико-синтетической обработки текста (источника знаний при использовании контекстологического словаря в качестве базы знаний или оригинала при использовании контекстологического словаря в качестве переводного словаря при переводе с одного языка на другой);
- пояснительный текст (или толкование, как в толковом словаре), который воспроизводит стереотипную ситуацию, предлагающую варианты образцов, стереотипных по отношению к тем, которые имеются в создавшейся реальной ситуации в обрабатываемом тексте, т. е. ситуации, которая инициировала поиск информации в контекстологическом словаре;
- генерирующий текст, который предлагает способы (обобщенные стереотипные примеры, термины, эквиваленты, их семантические связи) использования полученного нового знания для создания нового текста (например, перевода) или пути дальнейшего поиска, которые, как и в большин-

стве других словарно-справочных аппаратов, вводятся отсылочными пометами.

Являясь симбиозом базы знаний и толкового словаря, контекстологический словарь наиболее полно реализует свойство системности терминологии и наиболее наглядно изображает терминосистему конкретной предметной области. Эта концепция объясняет естественно-логическое разделение способов представления терминологии в контекстологическом словаре на два вида. Один из этих видов представляет основные (чаще родовые) термины, которые являются входными словами в структуре построения словаря и имеют подробное информационное описание. По второму способу представляются подчиненные (чаще видовые) термины, которые интерпретируются через определение основных терминов. Таким образом, создается естественная логическая сеть семантических отношений между терминами и воспроизводится структура терминосистемы предметной области, когда такое воспроизведение требуется для определения пути поиска в контекстологическом словаре по определенному информационному запросу.

Для ориентации в системных семантических отношениях терминов в целях практического использования в процессе поиска терминологической информации в контекстологическом словаре необходим оперативный инструмент поиска — некий адресный ключ. Функции такого адресного ключа может выполнять любой несложный поисковый аппарат, легко реализуемый при машинной интерпретации словаря. Те же функции может выполнять и обычный алфавитный указатель при традиционном оформлении контекстологического словаря полиграфическим способом. Кроме использования по своему прямому назначению, адресный ключ дает возможность судить об активности функционирования данного термина, например по количеству адресов. Эта информация извлекается из вне семантических отношений термина, в то время как анализ примеров контекстов дает достаточно полное представление именно о семантических отношениях внутри терминосистемы предметной области. С другой стороны, адресный ключ создает сеть оперативного поиска терминологической информации в контекстологическом словаре так же, как сеть поиска ин-

формации связывает фреймы в базе знаний. Если предлагаемый базой знаний фрейм слабо соответствует реальной ситуации, сеть поиска информации позволяет выбрать более подходящий фрейм путем прослеживания связей в системе семантически близких фреймов. Точно по такой же технологии может происходить взаимодействие пользователя (например, переводчика) с контекстологическим словарем.

Именно сеть, а не простые (даже и многосторонние) связи универсализирует контекстологический словарь и наделяет его функциями базы знаний, так как расширяет возможности словаря при формировании ответа на информационный запрос, т. е. при формировании нового образа, даже при недостаточном совпадении известного образа (например, для переводчика) с имеющимся в словаре. Адресный ключ, выполняющий в данном случае функции сети поиска терминологической информации, способен осуществлять мониторинг процесса расширения фактической информации о предлагаемом словарем образе, повышая наиболее важное в этом случае впечатление стереотипности, которое при достаточном накоплении знания по каналам семантических связей трансформируется в отношении качества в уверенное восприятие создаваемого образа и его экстраполирование на реальную ситуацию обрабатываемого текста (например, оригинала или перевода в зависимости от стадии аналитико-синтетического процесса).

На этом теоретическом фоне при переходе к практической интерпретации важным фактором подтверждения естественнологической системности терминологии предметной области является то, что входное слово, озаглавливающее словарную статью контекстологического словаря, объединяет несколько семантически связанных терминов, вводимых обобщенными примерами контекстов, и создает подсистему внутри терминосистемы. Таким образом, практически реализуется то, что происходит в базе знаний, когда связями объединяются семантически близкие фреймы в подсистему фреймов.

Именно это свойство контекстологического словаря наделяет его функциями базы знаний, так как такая подсистема, как правило, характеризуется логически полной оформленностью и способна сообщать логически завершенное и формально

оформленное знание чаще всего в виде формулировки и одновременно обобщенного примера имеющегося знания или модели для построения нового образа и нового знания.

Данные концепции были практически реализованы А. Л. Семеновым при построении контекстологического словаря предметной области маркетинга. Приведем пример статьи из этого словаря:

**144. Олигополистический рынок.** oligopolistic market.

Рынок, на котором небольшое количество *продавцов,* весьма чувствительных к политике *ценообразования* и *маркетинговым стратегиям* друг друга, торгуют с большим количеством *покупателей.*

A situation in which a few *sellers,* who are highly sensitive to each other's *pricing* and *marketing strategies,* sell to many buyers.

В данном примере проиллюстрированы все описанные выше особенности контекстологического словаря как организатора базы знаний.

## 3.6. Автоматический синтаксический анализ

Автоматический синтаксический анализ имеет целью с помощью алгоритмов получить в явном виде синтаксическую структуру предложения, простого и сложного (составного). При этом под «предложением» понимается чаще всего простое предложение, т. е. такое, в котором нет в качестве составляющих каких-либо других простых предложений, а сложное или составное предложение называют обычно «фразой».

Следует различать изображение синтаксической структуры, ее представление и обнаружение.

**Изображение синтаксической структуры**

Существуют разные способы изображения синтаксической структуры предложения, среди которых можно выделить три основных и наиболее распространенных: скобочная

запись; изображение зависимостей в виде стрелок, направленных от управляющего слова к управляемому; изображение синтаксической структуры в виде дерева.

Возьмем в качестве примера предложение

*Мальчик читает интересную книгу.*

- Скобочная запись структуры этого предложения будет иметь вид:

  *((Мальчик (читает) (интересную книгу))).*

  В данном случае в скобках заключены слова, непосредственно связанные друг с другом зависимостью синтаксического характера (определения связаны с определяемым словом, подлежащее связано со сказуемым. Отметим сразу, что без смыслового членения предложения и без обращения к семантике слов правильное определение синтаксической структуры невозможно, поэтому следует сразу отметить, что автоматический синтаксический анализ тесно связан и не может быть осуществлен без анализа семантической структуры предложения и его составляющих.

- Изображение синтаксической структуры в виде стрелок от управляющего слова к управляемому можно представить следующим образом:

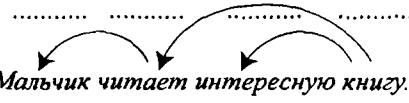


*Рис. 7.* **Синтаксическая структура предложения в виде стрелок**

- Запись в виде дерева предполагает изображение структуры предложения в виде некоторого графа:

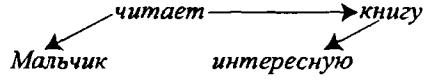


*Рис. 8.* **Синтаксическая структура предложения в виде дерева**

В каждом из способов изображения синтаксической структуры находит свое отражение способ представления синтаксической структуры.

**Представление синтаксической структуры**

В настоящее время можно отметить два основных способа представления синтаксической структуры, вариациями которых можно считать все другие способы. Таких вариаций может быть достаточно много, если учесть, что синтаксис, как сказано выше, тесно связан с семантикой, и при этом возникает много возможностей придумать разные системы отображения смысла в его связи с синтаксисом.

Если использовать термин «грамматика» для обозначения системы связей, то два наиболее популярных способа называются *«грамматикой непосредственно составляющих»* и *«грамматикой зависимостей»*.

Грамматика непосредственно составляющих отмечает в предложении наиболее связанные между собой по смыслу слова, и здесь показательным является скобочный способ изображения.

Грамматика зависимостей базируется на представлении об управлении, и в ее рамках наиболее популярен способ обозначения связей посредством стрелок. При этом возможны некоторые приближенные решения, поскольку управление, как известно из школьной грамматики, лишь один из видов связи (другие — согласование и примыкание).

Грамматика непосредственно составляющих построена на факте пространственного следования групп, составляющих предложение, — именных, глагольных, обстоятельственных и т. п. Здесь возможны разные виды связей. Предложение в целом рассматривается как совокупность групп, сводящихся к двум главным, образующим предложение, — группе имени (подлежащего) и группе глагола (сказуемого).

Имеется обширная литература, с разных точек зрения трактующая разные виды синтаксических зависимостей и способов их эксплицитного описания и алгоритмического обнаружения.

**Обнаружение синтаксической структуры**

В рамках компьютерной лингвистики разработаны различные способы автоматического обнаружения синтаксической структуры, ее эксплицированного описания на основе исследования структуры предложения естественного языка. Возможны разные основания классификации этих способов. А. Л. Василевский и Ю. Н. Марчук выделяют несколько групп из них в зависимости от оснований, по которым производится классификация. Можно исходить из двух оснований классификации: по способу движения по тексту в целях выявления структуры и по величине отрезков синтаксической структуры, которые при таком движении выделяются.

По способам движения по тексту выделяются методы *непрерывный* и *циклический.* При непрерывном методе движение по тексту осуществляется в одном направлении: слева направо (в большинстве действующих систем) или справа налево (реже — в зависимости от типа анализируемого языка). При этом за один просмотр текста выявляется, как правило, вся синтаксическая структура фразы в целом. Возможно также и определение синтаксической структуры составляющих фразу простых предложений. При этом под «предложением» понимается чаще всего простое предложение, т. е. такое, в котором нет в качестве составляющих каких-либо других простых предложений. Сложное или составное предложение обычно называют «фразой».

### 3.6.1. Современное состояние автоматического синтаксического анализа

В нашем пособии мы не будем слишком подробно останавливаться на автоматическом синтаксическом анализе. Достаточно обозначить основные подходы к практическому решению этой задачи. Дело в том, что как ранее, так и в настоящее время вокруг синтаксического анализа много бесплодных дискуссий, не имеющих никакого выхода в прагматику автоматического синтаксического анализа. Основанием и причиной такого рода положения является хорошо известная связь синтаксиса с семантикой. Она видна уже на уровне простых школь-

ных определений основных синтаксических категорий. Пример: что такое подлежащее? — то, о чем говорится в предложении, субъект действия. Что такое сказуемое? — то, что делает подлежащее.

Отсутствие формальных критериев определения синтаксических ролей слов в предложении приводит к тому, что создано множество всяких теорий вроде глубинных падежей, актуальных членений, ролевых структур и пр., которые, хотя и продвигают нас к определенному осмыслению состава и структуры предложения с точки зрения связей слов между собой, очень редко служат основанием действующих алгоритмов анализа и/или синтеза синтаксических структур. Те положения и рабочие теории, которые действительно служат рабочими основаниями, будут здесь рассмотрены.

Можно также сослаться на мнение ведущего специалиста по проблеме естественно-языкового интерфейса А. С. Нариньяни, который в (Диалог 95) говорит, что попытки построения такого интерфейса на основе синтаксического анализа делались в течение многих лет и не дали практических результатов, в то время как более простой и эффективный способ построения такого интерфейса заключается в учете семантики через лексику в рамках конкретных предметных областей и текстовых связей.

В настоящее время вновь возник интерес к автоматическому анализу и синтезу естественных языков у разработчиков-математиков и инженеров-программистов. Рассматривая анализатор текстов как некоторый конечный автомат с определенным числом состояний, они предлагают алгоритмы анализа, синтеза и перевода с естественного языка на другой (другие) естественный язык. Так, Д. Е. Шуклин полагает, что процесс общения с вычислительной системой на естественном языке можно условно разбить на операции разбора, анализа и синтеза. Операцию разбора можно определить как функцию преобразования текста на естественном языке из неформализованного вида в формализованное внутреннее представление, операцию анализа — как функцию преобразования данных, существующих во внутреннем представлении и вывода на их основе новых данных, также в формализованном виде. Опера-

цию синтеза можно представить как функцию формирования ответа на естественном языке, адекватного внутреннему формализованному представлению (Шуклин, 2002). Следует отметить, что подобный подход пока еще не увенчался практическим успехом. Трудности заключаются в переходе от высказывания на естественном языке к внутреннему представлению смысла (содержания) этого высказывания в некотором формализме. Однако новые возможности компьютерных технологий и результаты в формальном анализе высказываний, новые представления об искусственном интеллекте и другие достижения математической и компьютерной мысли заставляют относиться с вниманием к новым попыткам достижения практических результатов в этом направлении.

### 3.6.2. Синтаксическая структура

Отношения между словами в предложении определяются тем, что называют синтаксической структурой предложения. Выше мы рассмотрели синтаксические структуры словосочетаний, не обладающих предикативностью. Синтаксическая структура предложения должна включать в себя субъект, объект и предикат, а также все возможные их распространения.

Сначала введем исходные представления о фундаменте автоматического определения синтаксической структуры предложения.

*Изображение синтаксической структуры.* Существуют различные способы изображения синтаксической структуры предложения, среди которых можно выделить три: скобочная запись, изображение зависимостей в виде стрелок, направленных от управляющего слова к управляемому, и изображение синтаксической структуры в виде дерева (Прикладное языкознание, 1996; Василевский и др., 1970).

Примеры скобочной записи можно увидеть в современных работах по автоматическому анализу текста (например, в Constraints, 1994) и др. Запись в виде стрелок применяется в настоящее время чаще всего в учебных целях. Древесная форма записи довольно распространена.

*Представление синтаксической структуры.* Два основных способа представления синтаксической структуры существуют в настоящее время, вариациями которых можно считать все остальные способы. Если использовать термин «грамматика» для обозначения системы связей, то эти два способа называются «грамматика зависимостей» и «грамматика непосредственно составляющих». В первом случае синтаксическая структура рассматривается как некоторое дерево (граф), состоящее из ребер и узлов. В узлах графа помещаются слова предложения, ребра означают связи между словами. Эти связи отражают одну основную зависимость — управление. Некоторое слово в предложении управляет другими. При этом возможны некоторые приближения, поскольку управление, как известно, лишь один из видов связи (другие — согласование и примыкание). Тем не менее в грамматике зависимостей этот вид связи считается основным и главным.

Грамматика непосредственно составляющих построена на факте пространственного следования групп, составляющих предложение, — именных, глагольных и пр. Здесь возможны разные виды связей. Предложение в целом есть совокупность групп, сводящихся к двум главным, образующим предложение, — группе имени (подлежащего) и группе глагола (сказуемого).

Имеется обширная литература, трактующая разные способы представления синтаксических структур и связи между ними.

*Обнаружение синтаксической структуры.* Имеется также целая группа способов, которые позволяют обнаружить синтаксическую структуру предложения. Возможны разные основания классификации этих способов. А. Л. Василевский и Ю. Н. Марчук выделяют несколько групп из них в зависимости от оснований, по которым производится классификация. Можно исходить из двух оснований классификации: по способу движения по тексту в целях выявления структуры и по отрезкам синтаксической структуры, которые при таком движении выявляются.

По способам движения по тексту выделяются методы непрерывный и циклический. При непрерывном методе движение по тексту осуществляется в одном направлении: слева направо

(в большинстве случаев) или справа налево (реже). При этом за один просмотр текста выявляется, как правило, вся синтаксическая структура предложения в целом, при этом возможны также и варианты ее, некоторые из которых могут потом отсеиваться. Могут также определяться и какие-то составляющие синтаксической структуры предложения или фразы. Циклические методы заключаются в том, что предложение или фраза просматриваются повторно столько раз, сколько это необходимо: например, за первый просмотр выявляется подлежащее, за второй — сказуемое, далее определяются другие члены предложения. Цикличность заключается именно в этих повторных просмотрах текста, которые отсутствуют в непрерывных методах, где движение осуществляется однажды для всего анализа.

По объемам синтаксических структур, выявляемых при движении по тексту, способы анализа можно разделять на *интегральные* и *локальные*. Интегральные методы приводят к обнаружению всей структуры предложения сразу, локальные методы сводятся к установлению каких-то частей структуры. Можно в какой-то степени утверждать, что локальные методы связаны скорее всего с грамматикой непосредственно составляющих, а интегральные — с грамматикой зависимостей, хотя это и не обязательно.

### 3.6.3. Анализ по частям речи и членам предложения

Синтаксический анализ по частям речи и членам предложения является наиболее распространенным видом анализа в практически работающих системах. Основное его преимущество перед другими возможными видами синтаксического анализа и представления синтаксических структур можно сформулировать в виде следующего набора характеристик:

- анализ по частям речи и членам предложения отражает универсальную картину синтаксических связей в предложении, характеризующую большинство естественных языков;
- он содержит небольшое практически удобное число основных синтаксических категорий и связей слов в предложении;

- разрешающая сила этого вида анализа такова, что он обнаруживает основные свойства синтаксических связей для основного состава (основных классов) слов предложения и фразы;
- более тонкий синтаксический анализ может быть сделан после установления главных синтаксических связей и с опорой на них;
- наглядность и близость компонентов этого вида анализа к «человеческому» анализу предложения, знакомому из школьного образования языку, значительно облегчает последующее «человеческое» редактирование и корректировку результата.

Машинный результат в настоящее время практически всегда должен подвергаться корректировке со стороны человека-редактора, переводчика и/или пользователя системой. Если синтаксический анализ выполнен в категориях, незнакомых пользователю, то это вызывает дополнительные трудности в обработке текста и работе с компьютером.

Рассмотрим систему анализа по частям речи и членам предложения, принятую в системе машинного перевода АМПАР ВЦП (автоматизированный машинный перевод с английского языка на русский), которая (система анализа) осталась также в разработанных впоследствии на этой основе системах машинного перевода АНРАП, НЕРПА и СПРИНТ (Марчук 1983, Королев 1991 и др.).

Члены предложения суть универсалии, поскольку в каждом языке универсально представлены пять членов предложения. Когда переводчик производит перевод, сополагая текст с текстом, он осуществляет разбор предложения, рассматривая слова с точки зрения грамматической семантики. При этом семантика предложения членится им в терминах членов предложения. Все примеры на перевод даже весьма сложных случаев убедительно показывают, что переводчик не имеет никакой необходимости заниматься при переводе изучением механизмов словопроизводства, парадигматики или сиитаксиса в объеме, превышающем систему членов предложения. Даже перевод с китайского, в котором нет частей речи и соответствующей парадигматики, производится

таким образом, что переводчик выделяет в китайском предложении члены предложения и дальнейший анализ производит от них.

Таким образом, есть основания утверждать, что члены предложения являются универсальными семантическими эквивалентами в области грамматики, данными по тексту. Эти универсальные семантические эквиваленты могут служить удовлетворительной основой для организации сопоставительного анализа двух языков с целью нахождения переводных соответствий.

«В переводческом сопоставлении языки не выступают в качестве равноправных систем, каждая из которых должна сначала изучаться как самодовлеющее целое, а затем и сравниваться с другим языком в качестве единой системы. При сопоставлении, имеющем целью описание переводческих отношений, исходной точкой анализа служит язык оригинала. Задача анализа заключается в нахождении в текстах перевода отрезков, используемых для передачи значения единиц, которые выделяются в ИЯ, в изучаемых оригиналах. Таким образом, речь идет фактически не о сопоставлении систем двух языков, а об описании системы ИЯ в терминах системы ПЯ» (Комиссаров, 1973, с.197).

Часть речи может быть членом предложения или его частью. Часть речи как парадигматический класс входит в разные парадигмы. В некоторой парадигме часть речи может быть редуцированной формой члена предложения. Употребление части речи в данной парадигме означает редукцию парадигм парадигматического класса. Такое употребление части речи называют основным употреблением части речи (Ю. В. Рождественский, устное сообщение).

Идеальный текст представляет редуцированную форму членов предложения и редуцированную форму парадигматических классов, когда некоторая часть речи выступает в основном употреблении, а некоторый член предложения редуцирован до одного слова некоторой части речи. Эта корреляция между уровнями представляет собой использование свойства языка, которое формулируется так: максимальная единица нижележащего уровня равна минимальной единице вышележа-

щего уровня. Системная корреляция может быть осуществлена между всеми единицами данного и другого уровня, с наибольшим удобством между частями речи и членами предложения. Существует следующая схема корреляции между частями речи и членами предложения (табл. 1.).

*Таблица 1*

**Части речи — члены предложения**

| | Подлежащее | Сказуемое | Дополнение | Определение | Обстоятельство |
|---|---|---|---|---|---|
| Существительное | + | — | + | — | — |
| Прилагательное | — | + | — | + | — |
| Глагол | — | + | — | + | + |
| Наречие | — | — | — | — | + |
| Местоимение | + | | + | — | — |
| Числительное | + | — | + | + | — |

В данной таблице знак «+» означает обязательное наличие корреспонденции отдельного слова и члена предложения в соответствующем значении, данном в правом столбце матрицы. Знак «–» обозначает либо отсутствие такой корреспонденции, либо ее неуниверсальный характер, так как корреспонденции могут быть представлены не во всех языках (Рождественский, 1969).

Для целей автоматического анализа и синтеза целесообразно внести некоторые коррективы в традиционное представление о частях речи/членах предложения. Система частей речи должна учитывать такие классы, как местоимения, числительные, цифры, буквы и пр. Тогда как это делается в машинном переводе по принципу переводных соответствий, возникает классификация слов по лексико-грамматическим классам (с возможными семантическими и/или дистрибутивными подклассами), как это сделано в системе машинного перевода АМПАР и производных от нее системах.

В этих случаях неизбежно возникает синтаксическая (лексико-грамматическая) омонимия как принадлежность входной словоформы к одному из имеющихся в словаре синтаксических классов (подклассов) слов. Разрешение омонимии осуществляется соответствующими алгоритмами или системой алгоритмов, построенных по требованиям конкретной общей системы автоматической обработки текстов — например при использовании определенной системы стандартных операторов и пр.

### 3.6.4. Перспективы автоматического синтаксического анализа

За несколько десятков лет, в течение которых разрабатывались разные алгоритмы автоматического синтаксического анализа, не было получено ни одного, который бы удовлетворительно решал вопросы воссоздания синтактико-семантической структуры текста в формальных моделях. Наиболее приемлемым оказался синтаксический анализ в терминах частей речи и членов предложения. Действующие системы машинного перевода и другие автоматические системы анализа и синтеза текстов так или иначе воспроизводят именно этот способ анализа.

Это то, что касается глобальной постановки вопроса об автоматизации синтаксического анализа. В аспекте частных проблем автоматического синтаксического анализа имеется целый ряд достижений. Так, на этом пути есть решения синтаксического анализа (анализа отношений между членами группы слов) имен собственных, географических названий, именных групп с обозначением времени и числа и других локальных проблем. См., например, работы исследовательской группы университета г. Нанта по автоматическому анализу групп слов с обозначением дат, имен собственных и даже метафорических выражений (в части синтаксиса) (Dugas, 1993).

Из такого положения можно сделать вывод о том, что синтаксический анализ, будучи тесно связан с семантикой, поддается в настоящее время удовлетворительной формализации и алгоритмизации лишь в той мере, в какой поставленная прагматическая задача достаточно четко сформулирована и допус-

кает приближенное решение. Чаще всего таким требованиям удовлетворяют именно частные задачи, которые характеризуются наличием удовлетворительно работающей обратной связи, чего, как правило, нет в случае, если разрабатываются глобальные способы. Отрицательные результаты исследований по проекту ЕВРОТРА служат убедительным доказательством этому. Проект ЕВРОТРА — глобальный проект автоматического перевода со многих европейских языков на многие с помощью языка-посредника. Универсальный синтаксис языка-посредника оказался недостижимой задачей, и работы по проекту были фактически свернуты еще в середине восьмидесятых годов (Хроменков, 2000).

Тем не менее, следует отметить, что поиски более результативных методов и подходов должны быть продолжены. Возможно, что решения синтаксических проблем будут найдены на путях формализации синтаксического анализа одновременно с семантическим.

Последние исследования показали, что при создании естественно-языкового интерфейса, который является целью автоматического синтаксического анализа, чисто синтаксические соображения не всегда дают положительный результат. А. С. Нариньяни в обзорной статье по вопросам создания интерфейсов для работы с интеллектуальными системами говорит о том, что более двадцати лет абсолютное большинство групп, работающих в компьютерной лингвистике над решением проблемы естественно-языкового интерфейса (ЕЯ-интерфейса), пыталось решить эту проблему на основе различных вариантов синтаксически-ориентированного подхода. Фантастическое количество профессиональных усилий было затрачено на поиски наилучших грамматик для правил анализа и формализмов для представления результатов применения этих правил, но основная функция этих аппаратов остается все той же — построение синтаксической структуры входного предложения. С тем, чтобы уже после этого попытаться определить его смысл. Эта парадигма, говорит А. С. Нариньяни, оказалась абсолютно неадекватной для ЕЯ, поскольку именно понимание, а не синтаксическая структура, является главным для естественно-языкового интерфейса. Опыт многолетнего общения

профессионалов в области ЕЯ-интерфейса на разных языках позволил Нариньяни гораздо лучше понять психологию среднестатистического научного профессионала в этой области. Последнего на самом деле интересует не цель — практическое создание технологии ЕЯ-интерфейса. Подсознательно он даже не хочет этого, поскольку больше всего на свете ему нравится решать частные проблемы синтаксического анализа — области, с которой он психологически связан настолько тесно, что вопрос о неадекватности всего подхода просто не является для него релевантным (Диалог 95).

В середине 1990-х годов, говорит А. С. Нариньяни, мы могли убедиться в окончательном банкротстве традиционного подхода. Даже если в отдельных случаях некоторые профессиональные группы в состоянии ценой нескольких человеколет создать ЕЯ-интерфейс к какой-нибудь конкретной базе данных, то этому интерфейсу наверняка будут свойственны все те же врожденные дефекты ошибочно выбранной парадигмы. Любой синтаксически ориентированный ЕЯ-интерфейс будет дорогим, громоздким и сложным в настройке, а главное — весьма «хрупким» в отношении отклонений от нормы и чувствительным к выбору конкретной текстовой формы запроса.

Современная технология все больше базируется на семантически-ориентированном подходе. Тесная связь синтаксиса с семантикой не дает возможности создать эффективные системы и алгоритмы синтаксического анализа в отрыве от смысла высказывания и текста в целом. Представление о том, что формальный подход, базирующийся либо на анализе сверху вниз (от вершины предложения к его частям) или снизу вверх (от непосредственно составляющих низшего уровня до вершины предложения), даст возможность посредством сложения результатов получить полную и верную картину синтактико-семантических связей предложения, отвергается современными специалистами по компьютерной лингвистике. Достаточно назвать здесь работу известного специалиста по машинному переводу Ю. Цудзии. Он говорит, что структурные теории машинного перевода не только ничего не дали практике машинного перевода, но и завели ее в тупик вследствие того, что руководствовались мифом о так называемой композициональности (composi-

tionality) синтаксического анализа, когда принималось, что результаты анализа частей высказывания можно автоматически сложить в общую картину. Как утверждает Ю. Цудзии (Tsujii, 1997), перевод есть естественно-языковый процесс, синтаксический и семантический анализ не обладают таким свойством. Именно поэтому и говорится, что попытки создать естественно-языковый интерфейс на основе чисто формальных синтаксических методов уже двадцать лет не дают никаких результатов (см. Нариньян и выше).

Каков же выход из создавшегося положения?

По нашему мнению, которое подтверждается составом современных прикладных работ в области компьютерной лингвистики и машинного перевода, решение проблемы лежит в нахождении приемлемых практических решений в рамках конкретных языковых подсистем. Так, например, решаются вопросы машинного перевода текстов определенных микроподъязыков, в том числе и такого языка, как китайский (Зелко, 1991), рассматриваются теоретические и практические модели семантики подъязыков (Диалог 1995, 96, 97), создаются компьютерные модели получения семантических кластеров и групп слов на материале текстов определенных микроподъязыков (Arranz et al., 1997) и ведутся многочисленные другие работы в этом направлении. Имеются также эффективные методы выделения синтагм и синтаксических единств во входном тексте с помощью просодических, интонационных и других средств, характеризующих устную речь (Потапова, 2002). Можно утверждать, таким образом, что проблема автоматического синтаксического анализа решается в настоящее время индуктивным методом в рамках конкретных приложений и подъязыков.

## 3.7. Основные проблемы автоматического семантического анализа

Автоматический семантический анализ является одной из актуальных и вместе с тем наиболее сложной задачей компьютерной лингвистики. Нет необходимости доказывать, что он является необходимым этапом создания любой системы, моделирующей человеческий интеллект. Разработки проблем ав-то-

матического семантического анализа велись достаточно давно, однако каких-либо установившихся методов и приемов не было найдено. Можно утверждать, что имеющиеся решения практически пригодны только для ограниченных областей. В рамках конкретной предметной области можно выявить основной состав смысловых категорий, которые можно в исходном массиве текстов выделять автоматически. Поиски же универсальных путей автоматического анализа семантического содержания текстов пока не увенчались успехом, хотя работы в этом направлении ведутся. В самом начале исследований и разработок по машинному переводу были попытки придумать универсальные языки смысла, перевод на которые и с которых позволил бы сократить число бинарных алгоритмов. Выдвигались различные идеи организации таких универсальных языков. Однако ни одной действующей системы машинного перевода не было создано на основе универсальных языков смысла.

### 3.7.1. Автоматический семантический анализ

Обозначим некоторые принципиальные моменты, связанные с поставленными нами проблемами анализа на пути «слово–словосочетание–текст».

Под автоматическим семантическим анализом понимается совокупность методов и приемов, с помощью которых можно путем строгой и однозначной формальной процедуры, реализуемой на компьютере посредством специально разработанных лингвистических алгоритмов, с достаточной точностью представить смысл произвольного высказывания на естественном языке в виде последовательности символов, образующих некоторую формальную систему. Очевидно, что если были бы найдены удобные способы изображать в языке смысла любое произвольное высказывание на естественном языке, то вся остальная проблематика компьютерной лингвистики свелась бы к оптимальному кодированию, а в теоретическом плане, возможно, была бы решена проблема моделирования человеческого мышления. Однако формализация семантики, являющаяся необходимой предпосылкой такого умения выражать смысл, представляет собой крайне трудную задачу. Принципиальная сложность

проблемы состоит в том, что при изучении содержания (смысла) приходится выходить за пределы языка и обращаться к внешнему миру, к классификации предметов и представлений, лежащих вне сферы языка. Исследования в области формализации семантики, имеющие отношение к компьютерной лингвистике, можно условно разделить на два направления.

Первое направление включает в себя исследования, ведущиеся на дедуктивном абстрактно-теоретическом уровне и имеющие целью установить место семантики в рамках более общей науки о знаковых системах — семиотики и определить отношения между семантикой и другими составляющими семиотики — синтактикой и прагматикой, построить модели человеческого мышления самого по себе и в связи с процессом коммуникации; вывести универсальные закономерности образования понятий, связей между значениями слов и внутри высказывания между его составляющими; сформулировать и разрешить на уровне абстракции другие проблемы, относящиеся к человеческому мышлению и особенно к языковой деятельности человека в связи с мышлением и коммуникацией. Второе направление носит индуктивный, эмпирический характер. Оно ставит целью решение конкретных прикладных проблем, связанных с формализацией семантики, точнее, смысла в языковых выражениях, — проблем машинного перевода, автоматического информационного поиска, прикладных систем искусственного интеллекта. В рамках этого направления чаще всего рассматривается конкретная предметная область, система понятий в ограниченной области знания. Область формализации в таком случае представлена ограниченным семантическим полем. В тех случаях, когда имеет место такое ограничение предметной области, возможны вполне эффективные прикладные решения задач смыслового (семантического) анализа.

### 3.7.2. Смысл и текст

Из рассмотренного выше вытекает, что само понятие автоматического семантического анализа тесно связано с понятием «текст». Современное состояние автоматического семантического анализа отличается от результатов ранних исследо-

ваний только тем, что еще раз подчеркивается трудность получения данных о семантической структуре формализованным путем. Дело в том, что смысл, содержание сообщения неотделимы от того, что в настоящее время называется дискурсом. Содержание и форма неразрывны в том, что касается акта коммуникации, который происходит в совершенно конкретной обстановке и с учетом многих других факторов, в том числе и экстралингвистических. В. А. Звегинцев еще в 1976 году показал достаточно убедительно, что высказывание вне дискурса имеет не смысл, а псевдосмысл, и смысловое кодирование вне дискурса никакого смысла не имеет (Звегинцев, 1976). Перефразировки типа «Охотник ударом ноги убил волка», «Ударом ноги охотника волк был убит», «Волк был убит охотником посредством удара ногой» и пр. нельзя считать имеющими один и тот же смысл. Смысл определяется в конкретном дискурсе, и перестановка слов, замена словоформ, замена конструкций меняют и смысл, определяемый в зависимости от контекста, хотя с некоторой абстрактной точки зрения можно утверждать, что некоторый «глубинный» смысл всех высказываний одинаков. Дело в том и заключается, что «глубинность» и «одинаковость смысла» является некоторой обманной величиной — смысл не существует вне контекста.

Можно отметить, что теоретические работы по автоматическому синтаксическому и семантическому анализу не дали еще достаточно эффективных практических приложений и поэтому с некоторой дедуктивной точки зрения трудно судить о том, какие концепции в формализации смысла, содержания, плана содержания языка являются наиболее удобными для практических приложений. Имеется достаточное число формальных грамматик, описывающих механизмы соединения «единиц смысла» в высказывании. Однако они, как сказано выше, практически пригодны лишь для ограниченных приложений, хотя, как правило, их авторы постулируют универсальность предлагаемых подходов. А. С. Нариньяни замечает, что часто исследователи в области формальных грамматик с удовольствием занимаются частными вопросами формализации, даже не утруждая себя соображениями о том, насколько их системы подходят к анализу естественного языка и текстов в

целом. Естественно-языковый интерфейс, по его мнению, гораздо эффективнее строится на чисто семантической основе, выраженной главным образом через лексику. Поэтому слово как единица языка, словоформа как единица текста все более активно рассматриваются современными исследователями именно с точки зрения создания эффективных человеко-машинных интерфейсов (Диалог 1995, 96, 97).

Нам представляется заслуживающим внимания полностью привести некоторые высказывания А. С. Нариньяни по поводу синтаксической парадигмы. Он является одним из высококвалифицированных специалистов и долгое время занимался проблемами синтаксического и семантического анализа с точки зрения привыкшего мыслить абстрактно и обобщенно, как математик, а с другой стороны — как опытный программист, объективно оценивающий трудности алгоритмической и программной реализации. Так, в материалах конференции «Диалог-95» в статье «Проблема понимания ЕЯ-запросов к базам данных решена» (с. 206–215) он пишет следующее:

«Более двадцати лет абсолютное большинство групп, работающих в области компьютерной лингвистики над решением проблемы ЕЯ (естественно-языкового. — *Прим. Ю. Н. Марчука*) интерфейса, пытаются сделать это на основе различных вариантов синтаксически ориентированного подхода. Фантастическое количество профессиональных усилий было потрачено на поиски наилучших грамматик для правил анализа и формализмов для представления результатов применения этих правил, но основная функция этих аппаратов остается все той же — построение синтаксической структуры входного предложения. С тем, чтобы уже после его построения попытаться определить его смысл. (...)

Эта парадигма является абсолютно непригодной для ЕАИ (с. 207).

Состоялись многие десятки подобных дискуссий, чаще всего с демонстрацией наших программных систем, подтверждающих воплощение наших идей на практике. (...) Однако вся эта пропаганда никак не повлияла на поступательное движение величественного корабля компьютерной лингвистики, которое по-прежнему оставалось ориентированным на синтак-

сическую парадигму. (...) Данный опыт позволил гораздо лучше понять психологию среднестатистического научного профессионала, в частности, профессионала в области ЕАИ. На самом деле его не интересует цель — практическое создание технологии ЕАИ. Подсознательно он даже не хочет этого, поскольку больше всего на свете ему нравится решать частные проблемы синтаксического анализа, области, с которой он психологически связан настолько тесно, что вопрос о неадекватности всего подхода просто не является для него релевантным... Теперь, в середине 90-х годов, мы можем убедиться в окончательном банкротстве традиционного подхода».

Позитивная позиция А. С. Нариньяни заключается в провозглашении семантически ориентированного подхода. Она состоит в следующем: «Пытайся восстановить смысл текста, используя всю доступную семантическую и прагматическую информацию, обращайся к синтаксическим компонентам только тогда, когда это необходимо для разрешения неоднозначности; это обращение должно соответствовать требованию минимальной достаточности — используй минимум информации, необходимой для решения данной конкретной задачи» (Диалог 97, с. 208).

Приемы семантического анализа, связанные с такими представлениями, как семантические сети, фреймы и пр. рассмотрены нами в разделе 5 «Моделирование в компьютерной лингвистике информатике».

А. И. Новиков пишет: «Одной из основных трудностей изучения такого явления, как смысл, является его непосредственная ненаблюдаемость. Косвенным проявлением смысла, как известно, могут служить разного рода вторичные тексты: пересказ своими словами исходного текста, аннотация, реферат, конспект, наконец, представление содержания текста в виде набора ключевых слов, основных тезисов, планов и др. Деятельность по созданию вторичных текстов можно назвать вторичной текстовой деятельностью».

«Многообразие видов вторичных текстов определяется теми конкретными задачами, которые решает субъект вторичной текстовой деятельности и свидетельствует, как можно предположить, о неодномерности проекции текста. Результаты вто-

ричной деятельности можно рассматривать как обратную проекцию внутреннего представления текста вербальными средствами» (Новиков, 2002, с. 157). Далее А. И. Новиков предлагает некоторый аппарат и алгоритм выявления смысла текста, основанный на изучении действий информантов по извлечению «смысла» из текста.

При этом возникает различие между терминами «смысл» и «содержание» текста. А. И. Новиков описывает опыт установления такого различия. Информантам было дано три текста: один текст по химии с описанием химических реакций, второй текст исторический с описанием походов Александра Македонского и третий текст — художественный. В последнем описывается поведение двух людей — мужчины и женщины, которые едут в поезде и любуются из окна вагона проплывающим мимо пейзажем, обмениваясь замечаниями. Задание информантам: установить соответствие между «смыслом» и «содержанием» каждого текста. Результаты опроса информантов: в тексте по химии смысл полностью соответствует содержанию. В тексте по истории отмечена определенная разница между смыслом и содержанием: содержание — описание походов, смысл — показ завоевательной деятельности Александра Македонского. Третий текст характеризуется полным несовпадением смысла и содержания. Содержание — разговор, обмен замечаниями о природе. Смысл — выяснение отношений между мужчиной и женщиной.

Следует также отметить, что понятие «смысл» текста и «содержание» текста связаны с представлением о языковой картине мира. В разных естественных языках картины внешнего мира достаточно различны, что хорошо показано в современных исследованиях, посвященных картине мира и ее отражениям в лингвистике (см., например, Корнилов, 2003). В таких прикладных задачах, как перевод, различие в картинах мира проявляется наиболее ярко. Например в арабской ментальности понятие «верблюд» связано с представлением о стройности, красоте, оно целиком положительное. Но трудно себе представить, чтобы европейская женщина обрадовалась, если бы ей в порядке комплимента сказали, что она «красива как верблюд». Подобные проблемы приходится решать как в «че-

ловеческом», так и в машинном переводе. Однако и безотносительно к задаче перевода, например, в системе универсальных смысловых множителей нужно учитывать составляющие языковой картины мира.

Другим важным аспектом, влияющим на «смысл» и «содержание», является психология и психические конструкции. Этот вопрос подробно исследован в капитальном труде С. Э. Полякова «Мифы и реальность современной психологии» (Поляков 2004). Лингвистические аспекты современной психологии занимают в нем большое место, поскольку связь мышления и языка, а через мышление — языка и психики человеческой — в настоящее время требует эффективного компьютерного моделирования для построения удобного человеко-машинного интерфейса.

Таким образом, понятия смысла и содержания текстов трактуются по-разному в разных теориях. С прикладной точки зрения некоторые абстрактные категории, заложенные в «смысл», могут привести к приемлемым результатам, и попытки создать «языки смысла», «языки семантических множителей», с помощью которых можно описать некоторый конкретный набор ситуаций, в настоящее время ведутся. Так, в рамках некоторых структур Организации Объединенных Наций в течение определенного времени разрабатывался так называемый Universal Semantic Language — USL, который пока не получил широкого практического применения.

В любом случае можно утверждать, что основной лингвистической единицей, вокруг которой разрабатывается тот или иной прием определения смысла или содержания, является слово. Дело не только в том, что компьютер может легко идентифицировать слово (словоформу) как последовательность знаков между двумя соседними пробелами. Слово выступает в тексте как основной носитель содержания. Семантическое содержание современного слова не может рассматриваться без того, чтобы определить особенности его употребления в современных текстах. Поэтому мы рассмотрим слово как единицу предметного языка, языка определенной предметной области, какой оно и выступает в текстах.

Современный текст в большинстве случаев относится к тому, что мы называем деловой прозой или деловой речью. По крайней мере, компьютерная лингвистика чаще всего имеет дело именно с таким текстом. Главной же лексической составляющей деловой прозы является особое слово, которое называется «термином». Рассмотрим поэтому лингвистическую проблематику, связанную с термином и терминологией. Следует отметить, что многие учебники и материалы по компьютерной лингвистике не включают терминологию в состав этой науки, считая ее как бы периферийным явлением для лингвистических занятий. Между тем нельзя не согласиться с тем, что термин как слово современного языка не только в специальном общении, но и в широкой языковой коммуникации играет все большую роль. Представляется поэтому вполне оправданным включение основных положений науки о терминах в состав лексического раздела компьютерной лингвистики.

# *Глава 4*
# Терминология, терминоведение, терминография

## 4.1. Термин как лингвистическая проблема информатики

### 4.1.1. Терминология

Слово «терминология» многозначно, оно может означать как науку о терминах (хотя в последнее время укрепился термин «терминоведение», гораздо более точный в этом отношении), так и совокупность терминов конкретной области знания. Есть и другие значения этого слова. «Терминоведение» более адекватно описывает содержание науки о терминах. Терминография — наука о составлении терминологических словарей.

За последнее время появился ряд обобщающих работ по терминоведению, в которых рассматриваются также и собственно лингвистические аспекты терминов и терминологий как составляющих лексический уровень прикладной лингвистики. Можно отметить капитальный труд В. А. Татаринова «История отечественного терминоведения» (Татаринов В. А. История отечественного терминоведения. — М.: Московский Лицей, том 1, 1994. — 407 с.; том 2, 1995. — 333 с.), работы С. В. Гринева (Гринев С. В. Введение в терминоведение. — М.: Московский Лицей, 1993. — 309 с.), докторскую диссертацию С. Д. Шелова (Шелов С. Д. Опыт построения терминологической теории. — М.: МГУ. Дис... докт. филолог. наук, 1995. — 201 с.), монографию К. Я. Авербуха «Общая теория термина» (Иваново 2004, с. 249) и др.

Терминология как научная проблема и как составная часть словаря привлекает все большее внимание современного исследователя. В данном разделе нашей книги, руководствуясь замыслом и степенью актуальности лингвистических аспектов терминологии для информатики, рассмотрим следующее:

- роль терминологии и терминов в современном языке и в текстах, подлежащих автоматической переработке;
- проблемы анализа и синтеза терминологии и терминов;
- становление и развитие терминосистем;
- диахронический аспект современной терминологии и ее изучения;
- проблемы терминографии.

В современную эпоху информатизации подлежащие автоматической обработке тексты на естественном языке все в большей степени включают в себя научно-технические термины. Так, по данным ассоциации германских инженеров, в начале XX века терминология на немецком языке (вся научно-техническая терминология) насчитывала около 1 млн. слов. В середине века только терминология одной электротехники на немецком составляла уже около 4 млн. терминов (Better Translation 1983). В других языках рост числа терминов также является хорошо известным фактом.

### 4.1.2. Лексика современных текстов

Чем характеризуется современное состояние языка и речи — текстов — с точки зрения лексики? Во-первых, имеет место разделение этих текстов по стилю на тексты массовой коммуникации, литературные, художественные, научные, технические, рекламные, информационные и пр. Во всех этих текстах обязательно встречается лексика, общая для разных слоев этих текстов, т. е. имеет место функциональная, стилевая, авторская и пр. дифференциация лексики текстов; с этой точки зрения они не представляют собой однородной массы. Например, научные тексты содержат следующие разряды лексических единиц:

- слова общелитературного языка в значениях, принятых в литературном языке. Это предлоги, служебные слова, местоимения и т. п.;
- слова общелитературного языка, которые, как правило, в научном тексте имеют особое значение *(быть, иметь, мочь, часть, условие и пр.);*

- слова, характерные для научных текстов и лишь изредка употребляемые в ненаучных текстах *(анализировать, классифицировать, метод);*
- фразеологические выражения *(удовлетворить потребность, учесть замечание, взаимодействие компонентов);*
- специальная терминология данной отрасли науки *(мощность, множества, иероглиф, анод);*
- слова общелитературного языка, обычно не встречающиеся в научных текстах, но содержание которых может стать предметом научного рассмотрения *(абориген, абсурд, буран);*
- символы, гистограммы, схемы, математические выражения и т. п.

Ю. В. Рождественский, который проводит данную классификацию, дает следующее объяснение причин, по которым современная научная литература характеризуется определенным разнобоем в использовании лексики, происходит смешение стилей и возникают трудности в атрибуции лексики:

- появились комплексные научные исследования, представляющие собой объединение многих научных дисциплин, каждая из которых обладает своим предметом;
- составление отраслевых словарей, особенно словарей информатики, сталкивается с трудностью, выражающейся в том, что имеет место «смутное» состояние языка науки;
- планирование и управление наукой как особой отраслью деятельности связано с организацией документов и документооборота всех видов. Отсутствие нормы в области лексики научных текстов существенно ослабляет силу документов. Это положение объясняется неудовлетворительным состоянием языка и стиля науки. Если ранее словарь научных сочинений делился только на две категории: терминологию и общелитературную лексику, то теперь он делится на три части: общелитературную, общенаучную и специальную.

Развитие и изменение общенаучной лексики, составляющей основу научных текстов, идет быстро и хаотически. «Именно это развитие, как показывает анализ словарей, не

учитывается современным словарным хозяйством, в котором пока нет места специальному сбору, анализу и описанию общенаучной лексики. Общенаучная лексика ставит ряд проблем перед теорией лексикологии и лексикографии» (Теория и практика английской научной речи. — М.: МГУ, 1987. — 240 с., цит. с. 67).

Общие положения о характере современных научно-технических и общеязыковых текстов определяют характер современной терминологической работы. Хорошо известно, что терминологическая работа в масштабе страны имеет огромное значение для эффективности системы научно-технической информации. Последняя, в свою очередь, является фундаментом научно-технического прогресса. Каждое правительство, заинтересованное в совершенствовании инфраструктуры для научно-технического и социального развития, стремится совершенствовать терминологическую работу.

История отечественного терминоведения исчерпывающим образом изложена в капитальном труде В. А. Татаринова (Татаринов, 1994). От термина к терминоведению как науке, актуальной отрасли языкознания, — таков путь развития мысли. Но терминоведение не только языковедческая проблема. Значение терминологической работы определяется учеными-неязыковедами следующим образом:

- обеспечение взаимопонимания специалистов в рамках новейших пределов знания и родственных дисциплин;
- издание научно-технической литературы;
- эффективная подготовка научных и инженерных кадров;
- широкое применение ЭВМ в народном хозяйстве;
- информация и стандартизация документации;
- ведение работ по координации терминологии как внутри страны, так и в мировом масштабе (Сифоров и др., 1988).

Пафосом нормативного описания и регистрации термина является его фиксация в данной речи, описание в словарях и нормативных справочниках. Между тем с самого начала рассуждений о терминологии следует отметить ее динамичность, прямо противоречащую статическому нормативирующему подходу. Не говоря уже об общеупотребительной и общенауч-

ной лексике, которая служит как бы интерфейсом терминологической лексики с общенациональным языком, уже в пределах узкой терминологии имеет место четко выраженная динамика движения и развития терминов. «Как показывают последние исследования, любая отраслевая терминологическая система также предстает перед исследователем в виде метастабильной системы, находящейся лишь в относительно устойчивом состоянии. На любом этапе развития она в большей или меньшей степени охвачена движением (во всяком случае отдельные ее участки и пласты) и включает определенное количество динамических элементов» (Шевчук, 1988). Что касается самих этих динамических элементов, то здесь можно выделить либо сами термины, либо некоторые их характеристики — универбы, лексико-семантические варианты и пр.

Динамика развития отдельных терминосистем коррелирует с общенациональным языком и отражается не только в появлении новых слов и словосочетаний, но и в появлении новых значений у уже имеющихся слов.

## 4.2. Терминоведение

### 4.2.1. Наука о терминах

В «Лингвистическом энциклопедическом словаре» (Лингвистический энциклопедический словарь. — М.: Советская Энциклопедия, 1990. — 683 с.) слова «терминоведеиие» нет как отдельной статьи. Терминоведение упоминается только в статье «Термин» — наука о терминах. «Терминоведение» представляется более удачным как название науки, чем «терминология», поскольку последнее многозначно. Терминоведение появилось в 1969 году (Гринев, 1995). Терминоведение понимается как комплексная научно-прикладная дисциплина на стыке ряда областей науки. Основным объектом исследования в терминоведении являются специальные лексические единицы, в первую очередь термины. Основная цель терминоведения — изучение особенностей и закономерностей образования и развития терминологий для выработки рекомендаций по их совершенствованию и иаиболее эффективиому использованию. Для достижения этой

цели необходимо решение ряда теоретических и практических задач.

- К теоретическим задачам отиосятся:
  - установление и описание основных типов специальных лексических единиц, анализ особенностей, отличающих их от общеупотребительной лексики;
  - выработка общих методов описания и анализа терминологий;
  - определение общих свойств терминов и терминологий и особенностей их реализации;
  - исследование основных типов называемых терминами понятий и связей между ними;
  - изучение структурного и словообразовательного состава терминов;
  - исследование особенностей зарождения, образования и развития терминологий разных областей знания;
  - анализ особенностей функционирования терминов и терминологий в специальной речи, в научном общении;
  - совершенствование теоретических основ создания различных типов словарей специальной лексики и пр.
- К практическим задачам относятся следующие:
  - разработка методики нормализации (рекомендации и стандартизации) и создания терминологий в разных областях знания;
  - разработка методов терминологической работы;
  - установление критериев и принципов отбора и обработки специальной лексики;
  - разработка методов, приемов и рекомендаций по переводу терминов и пр.

### 4.2.2. Терминоведение и лингвистика

Терминоведение связано с рядом наук. В первую очередь оно имеет тесную связь с языкознанием, с его разделом — лексикологией, занимающейся изучением лексического состава языка. Терминоведение сформировалось в недрах лексикологии и использует практически все методы описания и анализа

общеупотребительной лексики. Важны для терминоведения также словообразование и синтаксис словосочетаний.

Тесная связь существует между терминоведением и научно-технической информацией. Некоторые информационные работы, например разработка средств лингвистического обеспечения информационных систем, носят преимущественно терминологический характер.

Роль специальной лексики в получении, хранении и передаче научных знаний, тесная связь истории специальной лексики с историей зарождения и развития научных понятий обусловливают связь терминоведения с гносеологией (теорией познания), науковедением и историей науки и техники. Эта связь усиливается в настоящее время в ходе работ по моделированию процессов приобретения специальных знаний и созданию систем искусственного интеллекта, в которых требуется такие знания представлять. Используются в терминоведении и различные математические методы, в первую очередь методы математической статистики.

С. В. Гринев выделяет ряд самостоятельных направлений в терминоведении. Так, общее терминоведение изучает наиболее общие свойства, процессы, происходящие в специальной лексике, а частное терминоведение занимается изучением отдельных областей знания. Семасиологическое терминоведение занимается исследованием проблем, связанных со значением (семантикой) специальных лексем, изменением значений и всевозможными семантическими явлениями — полисемией, омонимией и пр. Ономасиологическое терминоведение исследует структурные формы специальных лексем, процессы наименования специальных понятий и выбора оптимальных форм наименований. Историческое терминоведение изучает историю терминологий для того, чтобы вскрыть тенденции их образования и развития и с их учетом дать правильные рекомендации по их упорядочению.

В термине как лексической единице выделяются: содержательная структура, включающая значение и смысл, реализуемый в разных видах мотивированности; формальная структура, зависящая от наличия у термина языкового субстрата и реализуемая в виде фонетической, словообразовательной, словосочетатель-

ной формы вплоть до специфической терминологической; функциональная структура, включающая номинативную, сигнификативную, коммуникативную, прагматическую функции, а также эвристическую функцию, характерную только для терминов. На различных этапах формирования и функционирования специальных языков в их лексике существуют два основных вида объединения средств обозначения общих понятий: стихийно складывающиеся совокупности — терминологии, состоящие из предтерминов, и сознательно конструированные совокупности — терминосистемы, состоящие из терминов.

Все эти вопросы — содержание термина как лексической единицы, анализ терминологий и пр. — и входят в предмет терминоведения как лингвистической науки, касающейся других предметных областей.

Терминосистема представляет собой статичную модель теории, описывающей определенную область человеческих знаний как деятельности; соответственно, существует непременная взаимозависимость между отдельными терминами и теорией, элементами которой они являются.

Терминообразование — совокупность способов создания терминов, при которых образуются новые лексические единицы определенного естественного языка, имеющие специальное значение. Различается несколько видов терминообразования. Морфологический способ создания термина, или морфологическое терминообразование, — это способ создания термина-слова. В частности, терминопроизводство — это способ создания термина-производного слова, в котором аффикс имеет специальное значение, является терминоэлементом. Синтаксический способ создания термина, или синтаксическое терминообразование, — это способ создания термина-словосочетания.

Заимствование термина — это способ создания терминов, при котором лексические единицы переносятся из одного естественного языка в другой или же в язык специального общения.

### 4.2.3. Многозначность термина

Все эти и многие другие вопросы входят в состав терминоведения. Капитальная работа по теории терминоведения опубликована В. А. Татариновым в 1996 году — Теория тер-

миноведения (в 3 томах). Т.1. Теория термина: история и современное состояние. М.: Московский Лицей, 1996, 311 с. Подробный аналитический обзор истории становления лингвистической науки о терминах завершается (в первом томе) обзором теории термина в ее современном состоянии и терминоведческой типологией терминологической лексики, в которой весьма продуманно и обосновано решаются многие актуальные вопросы классификации лексики относительно терминов. В заключении этого тома обосновывается перспектива: разработка теории термина во всех ее аспектах, систематизация методов исследования терминологии, адекватное описание и внедрение результатов теории термина в прикладных исследованиях, создачие системы профессиональной подготовки терминологов.

Наиболее важной является разработка В. А. Татариновым проблемы многозначности термина. Структура значения термина в основном сохраняет особенности семантической структуры слова. Основной семантический объем термина задается его дефиницией и может варьироваться как интенсионально, так и экстенсионально в диахронии или в синхронии. Неоднозначность термина выдвигается на первый план, выступая в трех формах: амбисемии, эврисемии и полисемии.

Амбисемия характеризуется неопределенностью содержания, вызываемой не отсутствием точности в описании предмета мысли, а стремлением различных научных школ и отдельных исследователей к более глубокому проникновению в сущность объекта исследования, стремлением отразить его ранее не изученные стороны и аспекты.

Эврисемия отличает моносемный характер термина. Семантика термина-эврисеманта характеризуется высшей степенью обобщенности, что позволяет использовать его по отношению к неопределенному количеству денотатов (например, немецкий термин Gehaeuse — «корпус, оболочка»).

Полисемия — это способность термина иметь два или несколько взаимосвязанных значений, между которыми существуют отношения производности, взаимной мотивированности и категориальности. Полисемия термина выражает в определенной степени изоморфно отношения между логико-понятий-

ными категориями соответствующей сферы человеческой деятельности.

Проблема амбисемии, эврисемии и полисемии может быть охарактеризована через понятия интенсионала и экстенсионала. Термины-амбисеманты имеют неустойчивый интенсионал и чрезвычайно узкий экстенсионал. У эврисеманта — бесконечный экстенсионал при весьма узком интенсионале. Функционирование эврисемичных терминов приводит к постоянному расширению объема при сохранении его содержания. При полисемии происходит дробление интенсионала и расширение его экстенсионала.

Каждый разряд неоднозначных терминов имеет соответствующие языковые параметры проявления интенсиональных и экстенсиональных свойств. Амбисемия ярче всего прослеживается путем изучения дефинитивного представления терминов и их контекстуальных определений. Эврисемия может быть отслежена через максимальное фиксирование всех возможных словоупотреблений термина. Полисемия эксплицируется с помощью систематического воспроизведения лексического, синтаксического и тематического контекстов. Один дефиниционный анализ не в состоянии установить границы полисемии.

Языковые факты, пишет далее В. А. Татаринов, опровергают сложившееся мнение о стремлении термина к однозначности и необходимости моносемантизации всех терминов. Полисемия термина не является показателем его неточности. Чем сильнее развита система многозначности в терминологии, тем основательнее изучен предмет мысли, тем точнее установлены связи между общенаучными понятиями и отраслевым концептуальным аппаратом, тем категоризованнее и структурированнее предстает объект изучения. Полисемия термина отражает поступательный ход развития науки и выражает в семантической структуре слов принципы категоризации мира с помощью языка в не меньшей степени, чем это достигается путем отражения в морфологической структуре слов.

Констатация таких форм неоднозначности, как амбисемия и эврисемия, еще раз подтверждает, что не может быть знака равенства между понятиями точности (семантической опреде-

ленности) и однозначности термина. Это несравнимые понятия. Термин точен вне зависимости от принадлежности к тому или иному семантическому разряду (будь то амбисемичный, полисемичный или эврисемичный термин), но при условии, что объем выражаемого им понятия постоянно уточняется.

Такие семантические классы терминов, как синонимы, варианты, антонимы, партитивы, гипонимы, эквонимы и аспективы, имеют едва ли не большее распространение, чем соответствующие группы слов в общелитературном языке. Благодаря этим разрядам терминологических единиц выражаются многообразные логико-понятийные взаимосвязи в терминологии. С помощью множественности номинативных средств языка реализуется лабильность мыслительных структур.

Процессы терминообразования подчинены общим законам научно-мыслительной деятельности. Термины создаются как средство номинации понятий, предметов или явлений или как способ фиксации полученного знания. Для этого в терминообразовании сложились все необходимые средства: терминоэлементы, терминомодели и способы образования терминов (лексико-семантический, морфологический, синтаксический и способ заимствования). Значительное место в исследовании терминообразовательных процессов занимают понятия мотивированности, системности и внутренней формы термина.

Предметом изучения в теории термина являются также свойства термина, называемые когитальными свойствами. Это те свойства, в которых фокусируется логико-понятийное содержание мыслительных операций. Когитальные свойства термина примечательны тем, что они находят непременное выражение в терминологических структурах. Наибольшую гносеологическую ценность в терминологии имеют те когитальные свойства термина, которые приобрели категориальный статус.

## 4.3. Диахронические исследования в лексикографии и терминоведении

### 4.3.1. Диахронический аспект многозначности

Диахроническому аспекту развития многозначности слов основного терминологического слоя до последнего времени уделяется сравнительно мало внимания. Между тем здесь возникает возможность лучше понять характер номинации, уяснить, за счет каких резервов языка происходит расширение номенклатурных его возможностей, так необходимое и актуальное в современную эпоху. Одной из центральных проблем современного языкознания является становление и развитие предметных терминосистем в языках разных типов. При этом особый интерес представляют широкие терминосистемы, в рамках которых происходит взаимодействие общеупотребительной, общенаучной (общетехнической) и узкоспециальной лексики. На интуитивном уровне хорошо известно, что слова-термины тесно связаны со словами естественного языка. Они рождаются, живут и умирают так же, как и другие слова, которые мы называем общеупотребительными. Действительно, разве слова *«рука»*, *«нога»*, *«деньги»*, *«рынок»* — не термины? В то же время это слова общенародного языка. Поэтому «основной терминологический слой» — это слова, одновременно являющиеся терминами широкой предметной области и словами национального языка общеупотребительной сферы. Таких слов очень много в каждом языке.

Язык, особенно его лексический слой, — живая материя, он постоянно развивается. Слова приобретают новые значения, старые значения уходят, появляются также новые слова, все это совершается по законам языка, которые также требуют нового и постоянного изучения, поскольку они меняются тоже, хотя и более медленно. Динамика — это характеристика жизни термина как в диахронии, так и в синхронии: развитие многозначности, воздействие нормализации и социальных законов на язык, количественные характеристики распространенности, информационная нагрузка, роль в дискурсе и пр. На этом пути весьма актуально многоязычие, т. е. распространенность тер-

мина в разных языках, его современные переводы и смысловые эквиваленты в системах понятий (Марчук, 1996).

### 4.3.2. Диахронический вектор слова

В проведенном М. В. Марчук специальном исследовании диахронических проблем терминографии слов основного терминологического слоя показано, что в границах этого слоя можно выделить содержательным анализом и статистическими характеристиками некоторый представительный массив слов, у которых лексические значения меняются в течение определенного периода времени таким образом, что эти изменения можно обнаружить в словарях данного языка за этот период и в текстах. Был разработан аппарат для фиксации, измерения и наблюдения за диахроническими изменениями значений многозначных слов, названный диахроническим вектором слова, который представляет собой набор характеристик (число новых значений, число новых словосочетаний с данным словом, имеются или нет в словарной дефиниции перестановки переводов, какие значения ушли из имеющихся и какие появились новые, частотность слова и пр.). С помощью такого аппарата появилась возможность точными статистическими методами определить корреляцию между различными компонентами этого вектора и сделать содержательные выводы.

Как оказалось, наибольший прирост новых значений дает лексическая сочетаемость. Частотность сравнительно мало влияет на градиент изменения значений. Разные лексико-грамматические классы слов неодинаково характеризуются в аспекте изменения лексических значений, однако в целом слова выбранного слоя (с терминологическими значениями в области общей экономики) существенно изменились в составе и числе новых значений.

Слово «значение» применительно к лексическому значению, как было показано выше, может пониматься довольно широко. В данном исследовании значение определялось через перевод, что позволяло осуществлять как теоретическую оценку, так и перечислять практические импликации результатов исследования истории слова.

Идея диахронического вектора слова и количественно-качественных измерений значений может быть развита в другом, не менее отчетливом, как нам кажется, представлении.

## 4.4. Историческое развитие слова и его значений

### 4.4.1. Развитие терминологической лексики

Слово как основная составляющая языка (словарная единица) и текста (словоупотребление) развивается исторически. Прежде чем войти в современный словарь, оно проходит длительное развитие в языке и текстах на этом языке: основные его значения как бы кристаллизуются, выделяются через переводы. Текстоцентрический, лексицентрический и словарноцентрический подходы, из которых первый в настоящее время завоевывает как бы некоторый методический приоритет, соединяются вместе в составлении современного терминологического словаря. В самом этом терминологическом словаре лексика неоднородна: будучи плодом диахронического развития, с одной стороны, и современного состояния, синхронного среза нынешней языковой реальности — с другой, эта лексика представляет собой совокупность как некоторого основного ядра слов, базовых понятий, так и новых терминов, однозначных, появившихся недавно, отражающих новые, современные понятия.

### 4.4.2. Основной словарный состав

Нас особо интересует диахроническое развитие слов основного словарного состава, фундамента терминологии, на здании которого, за счет, в основном, словосочетаний, вырастает все терминологическое здание современного синхронного среза языка. Продолжим диахроническое исследование лексики основного терминологического слоя, начатое М. В. Марчук.

Будем вести исследование следующим образом:

- выделим некоторый массив исходных слов, слов-терминов, совпадающих со словами общелитературного языка, диа-

хроническое развитие которых в языке экономики проследим на основе имеющихся словарей;

- отберем словари, которые дали бы возможность на достаточно длительном и вместе с тем обозримом промежутке времени проследить развитие значений выбранных слов, а также словари, которые дали бы синхронную картину состояния экономической терминологии, желательно в некоторой языковой паре, чтобы выявить значение через перевод и тем придать такому явлению объективность;
- наметим некоторый формальный научно-исследовательский аппарат, с помощью которого можно было бы достаточно наглядно представить результаты исследования в части диахронического развития слова и его современного состояния;
- объективно исследуем полученную картину и сделаем выводы относительно закономерностей диахронического развития, связи синхронии и диахронии и конструкции терминологического поля в свете выбранной концепции.

### 4.4.3. Выбор слов исходного массива

Согласно идее М. В. Марчук, выберем подъязык экономики. Это весьма своеобразный подъязык естественного языка. Он содержит слова, безусловно, составляющие основу словарного фонда каждого языка, поскольку они обозначают фундаментальные понятия человеческого общества, социальной деятельности и самой жизни человека: деньги, товар, прибыль, капитал и пр. Эти слова, являясь терминами экономики, в то же самое время являются и словами общеупотребительного и общенаучного языка. История развития этих терминов — это история развития этих слов во всей их исторической ипостаси, от появления в языке до наших дней. Они являются фундаментом языка экономики, который строит систему экономических концептов вокруг этих слов, за счет словосочетаний и модификации значений.

Выберем следующий список английских слов для детального изучения:

| | |
|---|---|
| 1. capital | 3. deficit |
| 2. cost | 4. expense |
| 5. industry | 9. plant |
| 6. loan | 10. production |
| 7. market | 11. profit |
| 8. money | 12. trade |

Прокомментируем этот список.

Во-первых, отметим, что сюда включены существительные и рассматриваются только они. Это не означает, что автор не признает терминологического характера глаголов, прилагательных, наречий и пр. Существительные выбраны только потому, что в условиях ограниченной выборки изменения легче прослеживать у имен.

Во-вторых, список достаточно представительный с точки зрения содержания экономической науки. Наверное, можно было бы включить еще какие-то термины, однако и этих достаточно для обследования с целью получения общих зависимостей. Компактность списка гарантирует тщательность исследования отдельных слов.

Представляется, что определенную роль играет алфавитная равномерность списка, т. е. выбор слов по алфавиту такой, что более или менее равномерно покрывает объем словаря.

### 4.4.4. Выбор словарей

Выбор словарей определяется теми рассуждениями, которые уже сделаны, а именно:

- с диахронической точки зрения надо ориентироваться на достаточно представительный период времени;
- словари должны быть ориентированы на массового потребителя, а также включать такую лексику, которая составителями обычных словарей может рассматриваться как терминологическая;
- словари должны включать объяснения терминов, но в то же время не быть энциклопедиями, где дефиниции слишком подробны и сложны. Удобнее всего словари типа толковых;

- в синхронном плане, сопряженном с диахронным в нашей концепции, словарь должен быть переводным, т. е. давать значения слов в переводе на другой язык.

Исходя из этих соображений, были выбраны три английских толковых словаря и один англо-русский терминологический словарь по экономике.

Требуется определить диахронический срок, который отделяет один словарь от другого, — диахронический шаг исследования. Здесь нужно учесть следующее.

**Первое**. Какой интервал времени следует принять: постоянный или переменный.

В пользу постоянного говорят как будто бы статистические соображения, рассмотренные ранее. Действительно, например с позиций глоттохронологии, при обработке больших массивов данных за много веков существования письменности и человечества, применение точных статистических методов требует постоянного интервала времени, например 100 лет. Однако в нашу эпоху развитие происходит не равномерно, а скачкообразно, в том, например, смысле, что прибавление новых слов в абсолютном исчислении в любых языках мира подчиняется скорее экспоненциальному закону, чем линейному. Действительно, по данным лексикографов, прибавление вокабуляра, например, за счет технических терминов в течение XX века превосходит все пополнение за все предшествующие века. Это соответствует выводам ведущих футурологов (А. Тоффлер, например), которые утверждают, что развитие цивилизации можно измерять жизненным циклом одного поколения (30 лет), но за последний век это развитие идет гораздо более быстрыми абсолютными и относительными темпами, чем за все время истории человечества.

В связи с этим придерживаться постоянного временного интервала вряд ли целесообразно, принимая еще во внимание и то обстоятельство, что наши методы исследования скорее содержательные, чем формальные, они не включают точных статистических вычислений.

Более целесообразно ориентироваться на интервал одного поколения — 30 лет, тем более что словари, вышедшие в течение XX века, позволяют выдерживать такой интервал.

**Второе.** Каков должен быть объем словаря.

Чем подробнее и больше словарь, тем обширнее информация, которая может быть из него получена. Однако полные словари выходят достаточно редко, их трудно ориентировать на выбранный интервал. В крупных словарях чрезвычайно велико значение стиля словарного определения: при несовпадении стилей трудно анализировать даже сходные слова. Большой объем информации может оказаться ненужным для нашего исследования.

Словарь средних размеров — назовем так словари объема 70–100 тыс. слов — гораздо удобнее в работе. Такие словари ориентированы на массового и в то же время квалифицированного пользователя. (Подробный анализ словарей такого объема с точки зрения отбора лексики, структуры словарной статьи, соблюдения интересов пользователя и пр. представлен в интересной работе польских лексикографов Bilingual Lexicography in Poland: Theory and Practice. Ed. by Jan Wawrzynczyk, Warszawa, Univ. Warszawski,1995, — 137 p.). Сохраняя все необходимые характеристики всех выбранных слов, они в то же время дают их максимально сжато: требование краткости, таким образом, заставляет авторов выбрать самое главное в значении и дефиниции слова. Эта дефиниция становится юридически отработанной по краткости и содержательности. Словари такого рода обязательно дают сочетаемость слова, опять-таки основную, заслуживающую быть учтенной. Словари рассчитаны на перевод конкретных лексических единиц. Историзмы, диалектизмы, как правило, сопровождаются пометами.

**Третье.** В словарях одной фирмы, одной системы при переиздании часты повторения в определениях. Например, несколько подряд вышедших словарей Webster или Oxford разного времени могут иметь часть словарных дефиниций просто переписанных с предыдущего издания. В словарях разных авторов такого рода повторы встречаются реже. Поэтому мы выбрали для сравнения по возможности словари разных издательств.

**Четвертое.** Тип словаря. В соответствии со сказанным выше, три словаря выбираются толковых, один — терминоло-

гический переводной. Вследствие того, что поставленная задача требует характеристики происхождения слова хотя бы в минимальных размерах, например заимствовано оно или нет, важной характеристикой являлась бы этимологическая направленность словаря, хотя бы в небольшой степени. Это касается толковых словарей.

Требуется также обосновать выбор варианта английского языка — английский или американский, или оба. Было принято решение взять оба варианта в такой комбинации: исходный и заключительный толковые словари — английские, а промежуточный — американский.

С учетом указанных соображений были взяты следующие четыре словаря:

1. The Concise English Dictionary. Literary, Scientific and Technical. By Charles Annandale, Blackie & Son Limited, London, Glasgow and Bombay,1913, 848 pp.
2. The Concise Oxford Dictionary of Current English. Seventh edition. Ed. by J. B. Sykes. Oxford at the Clarendon Press, 1984, 1258 pp.
3. Webster's New Collegiate Dictionary. Springfield, Mass., USA, 1961, 1054 pp.
4. Англо-русский экономический словарь. Отв. ред. А. В. Аникин. М., Русский язык, 1981, 792 с.

## 4.5. Диахроническое дерево слова как инструмент исследования развития слова в диахронии

### 4.5.1. Диахроническое развитие слов

Чтобы лучше представить себе результаты исследования диахронического развития слова в духе предшествующего изложения, введем понятие «диахронического дерева слова». Назовем так конструкцию, представляющую истоки, промежуточное развитие и современное синхронное состояние слова, отраженного в виде такого дерева.

Диахроническое дерево представляет собой некоторый ориентированный граф, вершины которого представляют некоторое

состояние в определенный момент времени, а ребра — пути перехода в новое состояние за определенный период времени.

Назовем корнем диахронического дерева — КОР — точку, принадлежащую некоторому исходному состоянию дерева, отраженному в начальном из исследованных словарей. Точку, означающую состояние слова по данным следующего словаря, назовем ствол — СТВ. Наконец, точку, в которой описывается состояние слова в современном терминологическом словаре, назовем крона — КРО. Ребра, соединяющие эти точки, суть пути слова во времени. Они отражают развитие слова от одной точки, от одиого состояния к другому. Естественно, что на этих ребрах можно будет брать сколько угодно точек, каждая из которых может описываться соответствующим словарем.

В целом граф такого рода можно изобразить так, как это сделано на рис. 9:

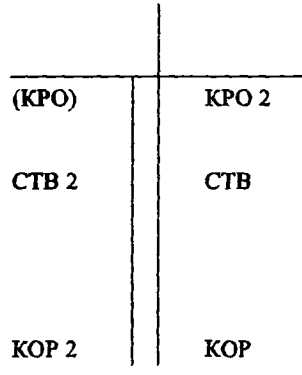


*Рис. 9.* **Диахроническое дерево слова**

Определим некоторые свойства точек и ребер графа.

Начнем с корня. Корень КОР может быть ординарным и множественным. Ординарный корень тот, у которого одно основное значение. Множественный корень имеет несколько основных значений. Фактически речь идет здесь о лексической омонимии. Можно было бы, вообще говоря, исключить лексическую омонимию, оставив только лексическую многозначность. Однако в таком случае мы потеряли бы часть диахрони-

ческой характеристики слова. Ординарный корень обозначается КОР, множественный — КОР 2.

При таком рассмотрении ствол дерева также может быть ординарным и множественным. Обозначения при этом аналогичны.

В отношении кроны отметим следующее. Вообще говоря, она отражает состояние слова по терминологическому словарю. Однако нередки случаи, когда терминологические сочетания даются уже в толковом словаре. Как правило, терминологический словарь имеет гораздо больший спектр этих сочетаний. Первичная крона, так сказать, подлежащая кроне терминологического словаря, будет обозначена КРО.

Естественно, что и на некоторых промежуточных точках между КОР, СТВ и КРО могут быть свои частные кроны, однако они в данном случае не отмечаются. Тем не менее, при подключении новых словарей они могут быть обозначены.

Ребра как таковые являются указателями периода времени и никакими особыми свойствами не обладают.

### 4.5.2. Направления движения по дереву

При исследовании слов с помощью дерева возможны в принципе два способа движения:

- от корня к кроне,
- от кроны к корню через ствол.

В практическом исследовании слов выборки мы будем использовать оба этих пути.

## 4.6. Исследование слов выборки с помощью диахронического дерева

### 4.6.1. Основная гипотеза

Прежде всего выскажем следующую гипотезу.

Для каждого слова выборки существует диахроническое дерево с полным набором компонент (с корнем, стволом и кроной).

Данное положение вовсе не является тривиальным. Мы убедились в этом, взяв некоторые слова, помимо слов выборки, а именно слова:

exuberance, dockage, message, privacy, prize, reside, thing.

При этом мы руководствовались наличием кроны, но не очень большой. Обнаружилось, что для слова privacy дерева фактически нет, поскольку значения, отмеченные в КОР и СТВ, не соответствуют значению кроны, которая к тому же очень незначительна, так что можно говорить вообще об отсутствии кроны.

Так что гипотеза о существовании имеет смысл.

### 4.6.2. Рассмотрение слов выборки

Рассмотрим теперь слова выборки, применяя понятие диахронического дерева.

Слово CAPITAL.

Следует отметить, что в отличие от других слов выборки здесь прилагательное и существительное даются в одной словарной статье во всех четырех словарях. Поэтому можно говорить, что на самом деле здесь как бы срослись два дерева: одно из них прилагательное, другое — существительное. В каждом из деревьев, образующих это двойное дерево, есть все «древесные» компоненты: корень, ствол, крона.

Несмотря на то что все выбранные словари разные, наблюдаем в англоязычных словарях сходство в толкованиях, что можно объяснить заимствованием. Например, в словаре Webster значение capital как adj объясняется как incurring the forefeiture of life, punishable with death. В словаре Annandale: incurring the forefeiture of life... etc. с помошью тех же слов. В словаре Oxford:... imploring loss of life; punishable by death. Далее имеет место практически полное совпадение всех значений прилагательного и существительного во всех трех англоязычных словарях.

Различия имеются в словосочетаниях. Словарь Annandale практически не выделяет каких-либо словосочетаний, описывая только семантику корня. Словарь Вебстера дает достаточно широкий перечень основных словосочетаний: capital account, capital expenditure, capital ship etc, которые не только включены в дефиннцию, но и вынесены отдельно. Словарь

Oxford дает практически все сочетания Вебстера и некоторые другие. Таким образом, можно сказать, что крона у этого слова начинается уже с середины ствола. КРО 1, подлежащая КРО, представлена достаточно густо.

Теперь рассмотрим КРО в словаре Аникина.

Прежде всего необходимо отметить, что здесь сохраняется побочная КРО от прилагательного — в порядке нумерации значений, т. е. существительное не отделено от прилагательного. Значения даны следующим образом:

Capital. 1. капитал, соц. фонды. 2. столица. 3. капитальный, основной, главный; самый важный.

Отметим, что значение capital punishment исчезло, видимо, вследствие экономической специфики словаря. Этой же спецификой объясняется большое количество терминологических словосочетаний. Сочетаний с глаголом отмечено 15, с прилагательным (типа agricultural capital) — 135, с существительными посредством предложного оборота (capital of a company) — 6. Среди сочетаний с прилагательным и причастием большое количество самых разнообразных вариантов значения и новых смыслов. Здесь есть также довольно большое количество словосочетаний, которые являются свободными, например constant capital, financial capital — *постоянный капитал*, *финансовый капитал* и пр. Включение таких словосочетаний в словарь можно объяснить стремлением создать определенные удобства пользователю, не слишком хорошо владеющему языком. Есть, однако, сочетания, перевод которых не может быть сделан без обращения к данному специальному словарю, например property capital — капитал в форме титулов собственности.

Таким образом, отмечаем, что для слова capital существует полное диахроническое дерево со всеми компонентами, с весьма развитой кроной в виде словосочетаний и некоторыми особенностями — двойным стволом, один их которых представлен прилагательным.

Рассмотрим следующее слово — COST.

Значений прилагательного здесь нет. Можно заметить, что последовательно в словарях Аннандейла, Вебстера и Оксфорд-

ском сохранились все основные значения этого слова. Интересно, что значение pain, suffering не то чтобы утратилось, но как-то постепенно стерлось: в Аннандейле мы имеем pain, suffering, в Вебстере сюда можно отнести значение под номером 2. Loss of any kind, detriment, в Оксфордском словаре вообще нет такого «гуманитарного» оттенка, слово cost стало чисто экономическим. Крона достаточно хорошо представлена уже в Оксфордском словаре главным образом за счет словосочетаний. Словарь Аникина имеет на это слово, пожалуй, наиболее крупную словарную статью, что и следовало ожидать.

Таким образом, мы отмечаем, что слово cost имеет все компоненты полного дерева. В некотором смысле крона «дегуманизировалась» в сторону чисто экономического эквивалента.

Следующее слово — DEFICIT.

Это слово имеет наименьшее количество значений из слов выборки. Самое краткое его определение дается в словаре Вебстера. В Оксфордском словаре уже появляется по крайней мере одно словосочетание, а именно deficit spending. При краткости определений в двух первых словарях они сильно отличаются друг от друга, равно как и от определения в Оксфордском словаре. Поэтому заимствование, видимо, исключено. В словаре Аникина это слово представлено большим количеством разнообразных словосочетаний: всего их 12, и они представляют собой разнообразные лексические единицы: budget deficit, external deficit, deficit of current account etc. Вывод: слово имеет полное диахроническое дерево, хорошо развитую крону с подкроной КРО 1, которая довольно бедна.

Слово EXPENSE.

Оно имеет в словарях Аннандейла и в Оксфордском по два основных значения: акт расходования и сумма, израсходованная. Интересно, что в словаре Вебстера, в котором есть еще и третье значение, а именно «источник расхода», в дефиниции вообще ни разу не упоминается слово «деньги» (money). Зато в последнем из английских словарей, а именно в Оксфордском, именно слово money пронизывает все определение. Здесь мы видим уже большое количество словосочетаний, образующих КРО.

В словаре Аникина, в отличие от всех других, слово EXPENSE употребляется только во множественном числе. Из всего этого можно сделать вывод: слово expense имеет полное диахроническое дерево, развитую крону, несколько отличающуюся от КРО1 множественным числом, и такой путь развития, который как бы приближает его к современным экономическим значениям: множественное число и слово money.

Слово INDUSTRY.

Это слово несколько своеобразно. В английских словарях Аннандейла и Оксфордском определения невелики, покрывают три значения: усердие в занятиях, промышленность, некоторое продуктивное занятие. Американский словарь дает четыре значения: кроме первых трех, с пометой «экономическое» (econ) дается: systematic labor or habitual employment, синоним — business. Конечно, это ответвление в значениях невелико, но можно утверждать, что в этом среднем месте у диахронического дерева появляется некоторое ответвление на середине ствола.

Кроны КРО1 практически нет, так как Оксфордский словарь не дает ни одного словосочетания, что достаточно редко. Зато словарь Аникина дает весьма «развесистую» крону. Вывод: диахроническое дерево также полносоставно.

Следующее слово — LOAN.

Оно несколько больше по объему, чем два предыдущих. Здесь интересно отметить, что уже в корне имеются некоторые ответвления, но не в части соседнего корня, как, например в слове capital, а представляющие некоторый подвид кроны: словосочетания loan-society, loan-office в словаре Аннандейла. Довольно значительное количество словосочетаний дает и словарь Вебстера. При сохранении основных значений в нем уже три словосочетания. В выходном же английском словаре Оксфорда значений четыре и все они связаны с новыми словосочетаниями (семь сочетаний разного экономического значения). В отношении этого слова, как и всех предыдущих, можно отметить, что новые значения слов имеют экономический характер, хотя в общем ничто не препятствует тому, чтобы эти

слова приобретали какой угодно оттенок. В словаре Аникина loan представлен большим количеством значений, главным образом за счет сочетаемости.

Таким образом, об этом слове можно сказать, что оно также имеет полное диахроническое дерево. Это дерево характеризуется небольшой кроной уже на уровне, близком к корню, а КРО имеет развитый характер.

Такая же в принципе картина, только гораздо более развитая, имеет место в слове MARKET. Здесь уже на уровне корня в словаре Аннандейла имеется 7 словосочетаний. Пять словосочетаний даются в словаре Вебстера (повторяют словарь Аннандейла). Все они носят экономический характер. Более двадцати словосочетаний в словаре Оксфорда. Здесь появляются такие слова, как Common market, stock-market, market-research и др., образующие богатую КРО1 в дереве. Словарь Аникина дает много словосочетаний. Развитие значений в этом слове идет за счет экономических значений, дерево полностью представлено.

Как и следовало ожидать, слово MONEY широко представлено уже в корне большим спектром словосочетаний. Здесь мы отметили 11 разных сочетаний на уровне КРО. При этом наблюдается единство определения основного значения — деньги, всеобщий эквивалент, бумажные деньги, ценные бумаги, эквиваленты денег. В выходном английском словаре Оксфорда уже имеют место 25 словосочетаний со словом money. Промежуточный словарь Вебстера практически сохраняет структуру Аннандейла, выделяя отдельные словосочетания в самостоятельные словарные статьи. Такая общая конструкция определений этого слова показывает древнюю роль денег, их исключительно важное место в развитии общества уже с момента появления. Интересно, что в словаре Аникина общее число словосочетаний со словом money не больше, чем у предыдущих слов, а даже меньше. Это, однако, не вносит корректив в основной вывод, а именно, что слово money представлено полносоставным диахроническим деревом с чрезвычайно богатой и древней кроной, начинающейся уже на уровне корня.

В слове PLANT переплелись на уровне корня два значения: одно из них — «растение», другое — «завод». Нас интересует второе. Это значение, которое у Аннандейла определено довольно скудно, зато четко, более подробно раскрыто в словаре Вебстера, причем внутри него выделены подзначения: сама фабрика (завод), оборудование института и пр. Кроны у корня нет, в средней части (словарь Вебстера) она также весьма незначительна (bicycle plant, скорее в порядке иллюстрации). КРО1 представлена всего одним сочетанием — plant-louse. В словаре Аникина имеется одна страница сочетаний с этим словом, причем можно отметить, что слово plant в значении «растение» также представлено в этом словаре, хотя и небольшой кроной.

Таким образом, и здесь имеется диахроническое дерево, причем с двумя корнями и стволами.

Слово PRODUCTION весьма невелико по объему словарной дефиниции во всех толковых словарях. Отмечены два значения: процесс производства и продукт производства. Это разделение последовательно проходит по всем трем словарям. В КРО можно включить только оксфордское словосочетание production line. В словаре Вебстера специально отмечается экономическое значение «создание экономической стоимости» (the creation of economic value), а также «производство товаров для удовлетворения человеческих потребностей». В словаре Аникина, напротив, сочетаниям с production уделено весьма много места, примерно столько же, сколько у слова money, хотя интуитивно можно предположить некоторую разницу между этими словами в терминах количества словосочетаний в пользу именно слова money. Можно отсюда сделать вывод о том, что количество стволов у дерева и величина его подкроны довольно слабо связаны с количеством ветвей кроны, хотя такая связь и имеется. Слово production, таким образом, имеет полное диахроническое дерево.

Исследование слов с точки зрения анализа словарной дефиниции для выявления структуры дерева приводит к вопросу, прямо не связанному с обозначенным подходом, но тем не менее достаточно интересному с точки зрения формального ана-

лиза. Связано ли количество слов в словарной дефиниции с общим значением, точнее, со значимостью слова в системе значений лексики данного языка или подъязыка? Очевидно, что величина словарной дефиниции в терминах количества слов безусловно связана с частотой слова: достаточно вспомнить, что наиболее обширные дефиниции характерны для частых слов вроде make, put (английский язык). Что же касается значимости слова и числа слов в дефиниции, то можно предположить, что зависимость имеется, хотя ее характер априори не ясен.

Переходя к очередному слову, а именно PROFIT, отмечаем, что словарная дефиниция его невелика по сравнению, например, с дефиницией слова money. Однозначно оно определяется в словаре Аннандейла, словарь Вебстера развивает исходные определения, выделяя слово во множественном числе (profits) для разграничения политэкономических категорий. Что касается кроны, то к ней можно отнести лишь rate of profit в словаре Аннандейла. Оксфордский словарь дает уже пять словосочетаний. Словарь Аникина уделяет слову полторы страницы словосочетаний экономического характера.

Заключительное слово — TRADE. По характеру определений оно представляет некоторую аналогию слову money, особенно в том, что касается наличия кроны у корня. Определение Аннандейла содержит уже большое количество словосочетаний с этим словом. Оксфордский словарь несколько увеличивает их число. При этом общая структура значений остается неизменной, равно как и их набор. Словарь Вебстера дает промежуточную картину. Для слова trade очевидна преемственность значений разных словарей. Примерно одинаков во всех трех словарях и состав словосочетаний. Характерно, что в словаре Аникина для этого слова, как и для слова money, всего одна страница словосочетаний. Полное дерево для слова trade безусловно существует.

Сравнивая результаты проверки дефиниций для обнаружения состава и структуры диахронического дерева слов, приходим к выводу, что деревья эти представлены разными типами структур. У некоторых из них имеется развитая крона уже у корня. Для таких деревьев, представителями которых являются

слова trade и money, последний словарь, терминологический переводной, дает относительно небольшое количество словосочетаний. Другие слова, такие как expense, plant, имеют корень без кроны, затем следует промежуточная крона у ствола, затем КРО1, как правило, хорошо отраженная в результирующей кроне. В этом случае можно говорить о том, что подкрона, отражающая состав общеупотребительных значений, соотносится с КРО как эти значения с терминологическими, т. е. можно рассуждать о соотношении этих типов значений. Другие типы деревьев представлены другими словами выборки.

Мы видим, таким образом, что существуют основания для типологии диахронических деревьев и что эта типология имеет предметное содержание, отражая особенности словоупотребления рассматриваемых слов и их отношение к другим словам и к терминологии в целом.

Так, мы можем обратить внимание на исходную многозначность слова. Для выявления этого фактора важны начальные состояния крон и стволов. Далее, можно поставить в центр исследования градиент, т. е. скорость приобретения словом новых значений. Эту характеристику можно поставить в зависимость от исходного состояния кроны. Наконец, можно учесть число значений разного типа в современном языке и связь каждого из них с этапами развития слов. В нашем исследовании важно то, что диахроническое дерево представляет собой удобный формальный аппарат для изучения направления развития значений, их числа, закономерностей прихода и появления новых значений — содержательный анализ находит свое выражение в формальных признаках и описаниях, что позволяет далее применять объективные формальные методы и алгоритмы.

Анализ такого рода особенно удобен для применения статистических, квантитативных подсчетов, на основе которых также можно будет сделать содержательные выводы, подобно тому, как это было сделано в кандидатской диссертации М. В. Марчук (Марчук, 1988) для выявления связей между компонентами диахронического вектора слова.

Диахронические характеристики чрезвычайно важны для всестороннего понимания нынешнего состояния слова в его разнообразных лексических коннотациях.

## 4.7. Анализ терминологии в современных информационных системах

### 4.7.1. Термины в информационных массивах

Начальные шаги на пути к прогрессу, например в развивающихся странах, связаны с созданием национальных терминологических систем — национальной терминологии уже установившихся областей, в которых часто преобладает европейская или американская терминология, и где надо создать собственные эквиваленты концептов.

Принципы применения компьютеров к анализу лексической информации были выработаны достаточно давно. См., например, диссертацию Е. В. Вертеля (Вертель, 1984). К настоящему времени имеются надежные способы анализа лексического состава самых разнообразных информатических текстов с помощью компьютерных методик. В настоящем параграфе мы выделим два направления. Первое: анализ терминологических систем с помощью словарей терминов, выявление общих закономерностей распределения терминологии в языках и в сферах коммуникации. Второе: изучение распределения самой терминологии в текстах, установление ее типов, систем концептов, устойчивости/неустойчивости терминологии в массивах текстов информационных систем. Первое направление как бы внешнее в отношении терминов, второе — внутреннее, оно затрагивает внутреннее функционирование терминов в информационных системах и массивах текстов.

Развитие терминологической лексикографии представляет собой интересный объект исследования. В нашей работе (Марчук, 1992) показано распределение терминологических словарей по разным странам, предметным областям и по времени. Последнее обстоятельство показывает динамический рост интереса к определенным областям науки или техники. Терминография — одновременно и следствие, и условие научно-технического прогресса. В терминологических словарях фиксируются материализованные компоненты научного знания. Внешняя картина развития словарно-терминологического дела позволяет оценить не только лексикографический прогресс, но и общую ситуацию развития науки и техники как в целом, так

и по отдельным составляющим (например, по странам и регионам мира).

### 4.7.2. Основные подходы в создании банков данных

Рассмотрим второе направление. Можно считать, что в части создания информационных систем и банков данных имеют место два основных подхода: Р. Ю. Кобрин обозначает их как а) построение языковых моделей предметных областей и б) построение алгебро-логических моделей баз данных (Кобрин, 1989). Можно сказать, что второе направление преобладает, поскольку современные базы данных создаются попрежнему не лингвистами, а специалистами в предметных областях и «информационщиками». Необязательно моделью информационной системы служат логико-алгебраические соображения, часто в основе систем лежат просто формализованные инженерные представления, основанные на здравом смысле, но тем не менее (а, может быть, именно потому) достаточно глубокие. Лингвистические же модели предметных областей распространены гораздо реже, и, чаще всего, они не касаются внутренней структуры модели, а выявляют через распределение лексики и терминов важные черты функционирования системы, но такие, которые занимают относительно периферийное положение. Пожалуй, лишь в работе Р. Ю. Кобрина лингвистическое моделирование поставлено в базис концептуального моделирования.

Более конкретно высказанные соображения можно проиллюстрировать разбором некоторых работ.

Наиболее важный вид научно-технической информации — это так называемая объектографическая информация, представляющая собой миожество описаний технических объектов — комплектующих изделий, материалов, технологий — состоящих из перечней конкретных признаков технических объектов (Хомутов 1989). Создание таких систем связано с решением двух основных проблем: первая заключается в определении способа структурного представления объектографической информации в виде, удобном для автоматизирован-

ной обработки, вторая — это разработка и внедрение программно-технологических комплексов, обеспечивающих функционирование систем.

Целью логического проектирования является создание общей схемы обработки информации, сохраняющей семантику информационных элементов при преобразовании данных в системе. Предметом логического проектирования служит логическая структура обработки информации, определяющая правила формального представления информации в виде данных и правила построения алгоритмов манипулирования ими. Основное условие построения систем такого типа заключается в том, что компоненты программно-технологического комплекса должны «выращиваться» (развиваться эволюционно) в связи с расширением и изменением информационных потребностей пользователей. Суть логического проектирования состоит в отображении упомянутой выше структуры требований к программно-технологическому комплексу на множество обеспечивающих средств системы.

Главной частью системы является модель предметной области, которая состоит из двух частей: прообраза предметной области и образа предметной области.

Прообраз предметной области составляют «знания об объектах», которые в общем случае состоят из описания свойств (характеристик) объектов, описания конкретных (явных) объектов и описания классов (неявных объектов). В свою очередь каждая характеристика представляет собой пару понятий: наименование и значение характеристики, которые задаются на входе системы парой конкретных терминов и образуют характеристику входного факта. Некоторая совокупность характеристик факта, поименованиая именем явного объекта, образует входное сообщение, которое в процессе ввода его в систему преобразуется в факты информациониой базы.

Каждый неявный объект представляет собой структурированное описание целого класса реальных объектов, обладающих одинаковой «номенклатурой» свойств и служит для задания информационной системы, с помощью которой отображаются объекты в памяти системы. Любой явный объект есть видовое понятие от соответствующего неявного объекта, полу-

чаемое из последнего присвоением уникального имени и уточнением характеристик неявного объекта их значениями.

Образом предметной области служит информационная база системы. Неявные объекты и входные сообщения непосредственно имеют соответствующие образы в структуре информационной базы.

Определяющим правилом выбора схемы образа предметной области является следующее: чем сложнее семантическая структура элемента информационной базы, тем меньше должен быть объем массива, реализующий его, и наоборот: чем больше может быть массив, реализующий некоторый элемент информационной базы, тем проще должна быть семантическая схема последнего.

Следствием этого является разделение всей информационной базы на словарные, информационные и технологические массивы.

В состав словарных массивов входят массивы терминологического, базового тематического и фактографического словарей, каждый из которых соответственно предназначен:

- терминологический — для отождествления внешнего представления лексических единиц с их внутренним представлением;
- базовый тематический — для фиксации, упорядочения и нормирования представления элементарных понятий, которые используются для описания в системе более сложных понятий;
- фактографический словарь — для регламентации описания объектов в системе, для нормирования лексики, в которой представляются характеристики объектов, и для адресации фактов в информационных массивах системы.

Непосредственными образами неявных объектов и входных сообщений соответственно являются тип объекта и факт информационной базы. Тип объекта — это элемент фактографического словаря, представляющий собой иерархию характеристик соответствующего неявного объекта. В фактографическом словаре возможно объединение типов объектов с помощью характеристик-отношений.

Факты информационной базы являются элементами информационных массивов, которые организуются по следующим правилам:

- общий справочный массив включает все факты по всем явным объектам, распознаваемым по словарям системы;
- специальные справочные массивы включают факты по всем явным объектам, но состав характеристик факта специально определяется фактографическим словарем для каждого массива;
- рабочие фактографические массивы включают факты по специально отмеченным в фактографическом словаре объектам и состав фактов каждого из них также определен в фактографическом словаре.

Эта структура в целом представляет собой, по мысли автора, «статическую» составляющую логической структуры предметной области системы. А ее «динамическая» составляющая представлена системой правил создания и ведения информационной базы и поиска фактов в массивах.

Правила поиска фактов задают способы реализации запросов к системе двух разных типов:

- тематический запрос — это запрос, в котором условия поиска формируются в терминах базового тематического словаря, отражающих имя типа объекта и наименования характеристик;
- предметный запрос — это запрос (по имени объекта, по значениям характеристик и по рекурсии), в котором условия поиска формулируются в терминах фактографического словаря. Таким образом, «статическая» и «динамическая» составляющие определяют все необходимые для работы системы аспекты формального представления информации в виде данных.

Из этого рассмотрения видно, что терминология, массивы терминов, словари терминологических областей представляют собой как бы вспомогательный материал, из которого строится система, а первичным являются информационно-логические и манипуляционные представления, исходя из которых строится общая структурная схема системы. Информационный процесс использует понятия «семантика», «логика», «понятие», на ко-

торых строятся концепты предметного и проблемного поля. Безусловно, в системе уделяется большое внимание анализу работы словарей, их полноте, корректировке в ходе работы и пр. Однако фундаментальный принцип — от логической информационной технологии к слову-выразителю понятий.

### 4.7.3. Лингвистический анализ терминологии

В отличие от такого подхода Р. Ю. Кобрин предлагает исходить при построении системы информационного типа из лингвистического анализа терминологии, от которого можно перейти к основам структуры предметного поля. Так, при создании банков картографических данных вне зависимости от конкретного содержания карты на ней присутствуют три вида информации: смысловая (семантическая) информация об объектах и характеристиках, выражаемая терминами естественного языка и системой условных знаков карты; метрическая информация, передающая пространственное положение объектов при помощи принятой системы координат; графическая информация о конфигурации, цвете, размерах условных знаков.

Для обработки этих видов информации разработан специальный лингвистический аппарат трансформации картографической информации в форму данных, обеспечивающий хранение, поиск и многоаспектную обработку данных на естественном языке (Кобрин, 1989). Информационно-технологическое обеспечение банка картографических данных представляет набор средств для идентификации, описания, структурирования и кодирования единиц информации, представленных в виде понятийно-терминологической системы предметной области и системы условных знаков, и ориентированных на автоматизацию процедур ввода, хранения и обработки картографической информации. В основу такого информационно-терминологического обеспечения положен принцип логико-лингвистического концептуального моделирования, обеспечивающий формализованное представление всех информационных единиц и отношений между ними в подсистемах, а именно: в базовой структуре, являющейся структурной моделью предметной области, представленной в виде графов и таблиц; в классификаторе картографической информации,

представляющем множество информационных единиц в выбранной системе кодирования.

В процессе конструирования информационно-терминологического обеспечения была создана терминологическая основа, в которой конкретизированы объекты предметной области в отношениях между ними в виде базового терминологического словника. Для отбора терминов использовался семантический критерий. Так, при разработке области ЛЕС создан словник, включающий 1500 терминов по лесоустроительной тематике и лесному картографированию. Анализ продуктивности терминов и их соответствия перечню базовых моделей, заранее определенных, позволил выделить класс базовых моделей терминов лесоустройства в виде, например таких:

П — С — «земельные угодья»,
С — С — «крутизна склона»,
С — П — С — «лес орехопромысловой зоны» и т. п.

Термины, образованные по этим моделям, включались в базовую структуру и в классификатор картографической информации и использовались в диалоге с ЭВМ. В словниках фиксировались основные парадигматические отношения.

Для формализации системы условных знаков были приняты особые модели и сформулированы определения основных информационных единиц, принятых в системе картографического объекта и картографической характеристики. Эти объекты выражаются не словами, а определенным набором полиграфических средств.

На картографических документах выделялись следующие типы смысловых отношений между информационными единицами: семантические, метрические, графические. Для их передачи был разработан специальный грамматический аппарат, заданный в базовой структуре, формализованный в классификаторе и предназначенный для передачи следующих видов отношений:

1) *род — вид*. Этот тип передает иерархические отношения;
2) *целое — часть*. Здесь передаются отношения включения / конкретизации, моделирующие структуры сложных объектов. Например объект «река» находится в отношении «целое — часть» к объектам «береговая линия»,«акватория»;

3) *объект — признак/свойство, характеристика.* Это наиболее распространенный тип отношения;
4) *действие — объект.* Отношение системное и реализуется программно. Возможны любые изменения объектов и характеристик в базе данных;
5) *объект — орудие действия.* Отношение системное и реализуется программно: возможно изменение характеристик при изменении объекта, изменение связанных объектов при изменении главного объекта и т. д.

Введены следующие грамматические средства: а) аппарат прерываний, б) аппарат связанных объектов, в) аппарат заполнений.

Аппарат прерываний — грамматическое средство, позволяющее структурировать сложный объект, т. е. представить все синтаксические связи объектов на стадии предмашинной подготовки и индексирования документов, а также формально реализовать эти связи на стадии ввода метрической информации. Например, у объекта «река» возможны прерывания объектами «береговая линия», «пристань», «маяк» и пр., что позволяет структурировать сложный объект «бассейн реки». Прерывание интерпретируется как реализация парадигматических отношений «целое — часть» и «объект — признак».

Аппарат связанных объектов — грамматическое средство, реализующее парадигматические отношения «целое — часть» и «объект — признак» в среде объектов, имеющих площадь (выражающихся в масштабе карты). Связанные объекты логически входят в структуру главного объекта, связаны с ним семантически, расположены на территории ведущего объекта. Например, у объекта «озеро» в качестве связанных могут выступать объекты «изобата», «остров», «паромная переправа» и т. д.

Аппарат заполнений реализует эти же типы отношений между «площадными» и дискретными (не имеющими площади, выражающейся в масштабе карты) объектами. Объекты и характеристики заполнения обязательно расположены в границах главного объекта и связаны с ним логически. Для объекта «озеро» заполнениями являются дискретные объекты «характеристика грунта», «отметка глубины» и пр.

Грамматические средства реализованы программно и закреплены в структурах интегрального файла, имеющего спе-

циальные поля прерывания, связности и заполнений. В форматах базовой структуры перечислены информационные единицы, находящиеся в фиксированных в концептуальной модели парадигматических отношениях. Эти единицы сгруппированы в списки классификатора и предъявляются пользователю в процессе индексирования. Синтагматические отношения могут устанавливаться между произвольно выбранными из классификатора информационными единицами в процессе индексирования и не задаются в базовой структуре.

Базовая структура позволяет строить фреймы, описывающие конкретные достаточно сложные ситуации. При этом возможно получение контекстуальных знаний, не содержащихся непосредственно в знаках исходного документа. Например, при обработке раздела карты «Гидрография и гидротехнические сооружения» возможно структурировать фреймы «бассейн реки», «берега, удобные для строительства», «фарватеры большегрузных рек» и т. д., непосредственно в знаках картографии не выраженные. Созданные на базе данной концепции системы ориентированы на любые виды картографических документов.

### 4.7.4. Построение системы логических отношений

Таким образом, мы видим, что посредством изучения состава, структуры и распределения терминов в текстах документов можно построить систему логических (семантических) отношений в рамках заданной предметной области. Путь создания проходит не от априорных логико-семантических отношений, как это было в первом из рассмотренных подходов, а через установление отношений путем изучения их проявлений в лексике, в терминологических словосочетаниях, путем изучения соответствующих ситуационных фреймов.

Как показывает современный опыт, такой путь наиболее эффективен в создании человеко-машинных интерфейсов. Если априорно-логическое построение требует некоторой заранее заданной картины ситуации в предметной области и системы отношений объектов, то изучение посредством анализа распределения терминов дает возможность индуктивным путем прийти к построению объективно существующей картины предмет-

ной области в данный момент. Эта картина может существенно отличаться от заранее постулированной дедуктивно картины. Здесь также может оказаться весьма полезным исследование статистического распределения терминов в текстах.

### 4.7.5. Статистическое распределение терминов

Изучение статистического распределения терминов в текстах может дать весьма интересные результаты. Современное состояние исследований статистических характеристик текста идет по нескольким основным направлениям. В данном месте мы рассмотрим некоторые общие идеи, высказанные относительно качеств текста и вытекающие из глобального исследования статистики терминов.

Для активного использования научного опыта, отраженного в информационных массивах систем, недостаточно только получать новую научную информацию, например о соответствующих разработках, чтобы потом повторить их или каким-либо образом использовать в собственных исследованиях. Гораздо важнее и эффективнее выявлять в информационных массивах не конкретные результаты, а основные новые тенденции и направления развития науки. Тогда работа в новых, перспективных направлениях и соответствующий поиск могут дать гораздо больше, чем простое заимствование результатов.

Поскольку основные процессы и результаты научно-исследовательской деятельности находят отражение в информационных документальных массивах, то, используя возможности современной информационной технологии для их обработки, можно найти закономерности в динамике поведения элементов информации баз данных и на их основе разработать способы определения новых тенденций.

Показательной в этом отношении является работа, посвященная анализу роли новых терминов в появлении нового знания. Е. Ю. Павловска, работавшая в Болгарии, на материале текстов по медицине и биологии высказала предположение, что «поведение» лексики, используемой для формирования текстов документов в базе данных, обладает определенной динамикой во времени и позволяет обнаружить некоторые зако-

номерности. Эти закономерности проявляются в том, что при возникновении нового направления в научных исследованиях статистические характеристики некоторых терминов, используемых в описаниях и текстах документов, обладают повышенной нестабильностью. Частота встречаемости этих терминов в базах данных носит колебательный характер. Это связано с неустановившимся процессом формирования терминов — ключевых слов, отражающих основной смысл публикаций в данной области. Через определенное время эти колебания прекращаются и частота появления в прошлом «колеблющегося» термина начинает расти во времени, либо падать, либо остается постоянной величиной.

В работе (Павловска, 1988) доказано, что лексический состав проблемно ориентированной базы данных является открытой системой, находящейся в неравновесном состоянии. Появление порядка в такой системе возможно только при наличии внешних потоков (вещественно-энергетических или информационных). В данном случае роль внешних возмущений играют новые термины, вводимые авторами научных публикаций в систему терминологии.

Развитие такой сложной динамической системы, как лексический состав проблемно-ориентированной базы данных, может быть представлен следующим образом: на «спокойном» этапе развития данной научной дисциплины система терминологии устойчива. Она ликвидирует любые отклонения и снова возвращается в это состояние. Но со временем в результате постоянного изменения соответствующих характеристик устойчивость терминологической системы постепенно падает — «ослабевает». Система превращается в неустойчивую. Количество неподавляемых возмущений резко возрастает, и система переходит в новое устойчивое состояние.

При возникновении нового направления исследований, новой научной дисциплины интенсивный колебательный рост частоты встречаемости определенных терминов, связанный с ростом количества публикаций по данной проблеме, резко нарушает упорядоченность лексического состава проблемно-ориентированной базы данных. Энтропия системы лексики возрастает. Система становится неустойчивой, но рост энтропии в данном случае

обеспечивает системе возможность эволюционного перехода в новое устойчивое состояние с меньшей энтропией.

Система терминологии носит диссипативный характер, так как прекращение соответствующих внешних потоков (новых терминов, обозначающих новые идеи, направления, методы исследования), питающих и поддерживающих структуру, приводит ее к разрушению — диссипации. Применительно к поставленной задаче можно сказать, что неустойчивая система, неустойчивое состояние лексики в проблемно-ориентированной базе данных может рассматриваться как сигнал о начале развития новых направлений в исследовании.

Метод, разработанный автором для определения наличия нового направления, включает несколько процедур:

- выделение в базе данных проблемной подбазы, содержащей документы в выбранной тематической области;
- выбор временного интервала исследования и группировка документов в тематической подбазе по хронологическому принципу;
- формирование стоп-словаря, предназначенного для исключения неинформативной лексики и автоматическое составление словаря специальных терминов, характерных для данной тематической области;
- автоматическое выявление словосочетаний с помощью созданных частотных словарей однословных терминов. Эта процедура важна потому, что в заглавиях документов словосочетания присутствуют в неявной форме;
- обработка полученных частотных словарей однословных терминов и словосочетаний, заключающаяся в нормировании частот, пороговом выделении информативных терминов и вычислении трендов (тенденций) — определении характера поведения лексических единиц базы данных во времени;
- формирование перечня однословных терминов и словосочетаний, которые предоставляются потребителям информации. Поскольку эти термины, в соответствии с описанной моделью, характеризуют начало развития нового направления исследований, то их включение в поисковый образ запроса должно привести к извлечению из базы дан-

ных документов, содержащих «новое» или еще «неизвестное» по данной проблеме.

Предложенная методика прошла проверку на массивах медицинской литературы по теме «моноклональные антитела». Полученные списки терминов и словосочетаний были переданы на экспертизу специалистов с целью:

- определения, в какой степени однословные термины и словосочетания, полученные с помощью предложенного метода, являются сигналом формирования новых тенденций в исследованиях в данной тематической области;
- оценки грамматической правильности (истинности) словосочетаний, полученных автоматизированным путем.

В качестве экспертов были приглашены ведущие ученые Болгарии в области иммунологии, онкологии и гибридомной технологии, т. е. в тех областях медицинской биотехнологии, в которых моноклональные антитела находят в последние годы наибольшее применение. Оценивая списки, эксперты отметили, что наибольший интерес представляют перечни однословных терминов и словосочетаний, имевших колебательный характер в 1981–85 годы. Большая часть терминов отражает наиболее перспективные в настоящее время направления в области медицинской биотехнологии, производства и использования моноклональных антител, в отличие от терминов и словосочетаний с возрастающей частотой встречаемости, которые в своем большинстве свидетельствуют о более широком развитии и применении известных несколько ранее методов и направлений в исследованиях. Количество терминов и словосочетаний с непостоянными частотными характеристиками, описывающих новые пути в исследованиях, новые методики, новые направления в тематической области «моноклональные антитела», появившиеся в начале 1980-х годов, а к настоящему моменту утвердившиеся, ставшие актуальными, превышает количество подобных терминов в остальных списках в 2,5 раза.

В списках словосочетаний количество «ложных» словосочетаний находилось в пределах от 8,8 до 13%. Метод выделения словосочетаний обладает хорошими характеристиками, поскольку количество истинных словосочетаний для всех трех групп составило 88,7 % от общего количества комбинаций тер-

минов. Эксперты также отметили, что для ориентации исследователей на перспективные направления работ в их узкой тематической области метод анализа словосочетаний более эффективен, чем анализ однословных терминов.

Заключительным этапом эксперимента явился информационный поиск в базе данных по запросам, основными элементами которых были выделенные однословные термины и словосочетания, «поведение» которых удовлетворяло условиям методики, т. е. термины, статистические характеристики которых определяли колебательный характер частот появления этих терминов в документах или их рост. Цель поиска — убедиться, действительно ли с помощью полученных слов и словосочетаний можно отыскивать в базе данных наиболее актуальную информацию по конкретным проблемам, т. е. проверить работоспособность метода. Эксперимент, в ходе которого было проведено 55 поисков, дал положительные результаты (Павловска 1988).

Таким образом, мы видим, что возможности исследования распределений терминов в словарях и текстах, возможности использования терминологических словарей в сочетании с применением ЭВМ на больших массивах словарных баз данных и текстов расширяются и приводят к новым результатам. Появляется реальная возможность увеличить объективность и повысить качество исследования.

## 4.8. Терминологические словосочетания

### 4.8.1. Общие представления о терминологических словосочетаниях

Проблема словосочетания является одной из главных в языкознании и не уступает по сложности и количеству различных постановок проблеме слова. Мы рассмотрим ее с точки зрения прикладного языкознания и информатики. О лингвистической трактовке словосочетаний мы говорили в разделе 3.4.

Языкознание различает устойчивые, традиционно повторяющиеся сочетания слов и сочетания переменные, свободно создаваемые в процессе речи (Маслов, 1987). Несмотря на то,

что точной и общепринятой классификации словосочетаний нет, в компьютерной лингвистике они играют большую роль, как это видно на примерах, описанных в предыдущем разделе относительно роли новой терминологии в квалификации текстов. Во многих анализирующих лексику системах автоматической обработки текстов найдены эффективные пути анализа и синтеза словосочетаний.

Словосочетания играют огромную роль в передаче смысла. Так, в предыдущем разделе было показано, насколько актуальны не только отдельно стоящие термины, но и особенно сочетания терминов, в установлении нового содержания текста. В терминологических словарях всякого рода словосочетания занимают ведущее место по сравнению с отдельно употребленными терминами. В автоматизированных и автоматических словарях словосочетания составляют до 50 % объема и даже более. Поэтому роль анализа словосочетаний в информатике чрезвычайно велика.

Рассмотрим сначала роль программного обеспечения и возможностей программ в обработке словосочетаний естественного языка. Собственно языки программирования в этой области достигли пока немногого, но работы продолжаются и перспективы представляются положительными (Брябрин, 1988). Для массового пользователя разрабатываются так называемые пакеты прикладных программ для решения наиболее массовых типовых задач в самых различных сферах применения — от программ составления и редактирования деловых писем до статистической обработки массивов данных. Наличие большого числа прикладных программ избавляет пользователя от программирования, а выбор необходимой программы и работа с ней обеспечиваются диалоговой системой, в основе которой лежит искусственный язык. Такой подход достаточно продуктивен. Однако, если выявить общие положения, лежащие в его основе, то представится, что они базируются на существенных ограничениях естественного языка. Так, упор делается на формальную обработку уже освоенного уровня языка — уровня отдельных слов и включение его в структуру естественного языка. Сведение многословных наименований понятий к однословным позволяет не выходить за рамки освоенно-

го уровня языка. Это ведет к появлению естественно-искусственных слов, составленных из начал слов, ограниченных по длине (Лингвистическая прагматика, 1989).

Благодаря ориентации на уровень отдельных слов практически решен вопрос о пределах редукции естественного языка, в первую очередь вопросительных и повелительных предложений. И в вопросах, и в командах оказалось возможным ограничиться только именами. Построение запросов и команд в форме цепочки имен, естественных или искусственных, связанных знаками искусственного синтаксиса, сохранило преемственность конструкций языков общения с конструкциями языков программирования. Одновременно это означает начало конструирования естественно-искусственных языков для человеко-машинных коммуникаций. Однако жесткий синтаксис языков программирования и состав используемых знаков синтаксиса все еще представляет существенные неудобства для конечного пользователя. К этому можно добавить несовпадение логических структур искусственных и естественных языков.

Арсенал языковых средств промышленных информационных систем документального типа преимущественно использует уровень отдельных слов — хорошо известные языки дескрипторного типа с грамматикой или без нее. Этот же уровень использования естественного языка имеет место и в промышленных фактографических информационных системах и базах данных, где таким образом строятся языки доступа и работы с данными. В некоторых последних моделях ЭВМ используется принципиально иной путь построения средств общения для массового конечного пользователя — последовательные выборы на основе системы меню.

В пределах имеющихся прикладных программ язык имен или пиктограмм, варианты которого реализованы в языках типа меню, по существу, не требует от человека применения вербальных средств общения. Однако определенные успехи этих и подобных им подходов инженерной природы в создании диалоговых средств общения с ЭВМ не дают оснований для утверждения о решении проблемы общения. Для нетиповых приложений, а они всегда проявляются в профессиональной деятельности человека, и, можно добавить, всегда наиболее

важны и перспективны, языковые средства современных диалоговых систем слишком ограничены. Необходимость их совершенствования остается главной задачей. В первую очередь необходимо расширение естественно-языковой компоненты в искусственных языках: переход на уровень словосочетаний, снятие ограничений на длину имен, упрощение синтаксических конструкций в смысле приближения их логической структуры к конструкциям естественного языка, а также выбор удобных для запоминания символов.

Естественный язык на уровне предложений и на уровне слов и словосочетаний — две части одной и той же проблемы. Они являются, по существу, последовательными этапами решения одной проблемы. Но без успешного освоения уровня словосочетаний трудно надеяться на практическое использование уровня предложений и более крупных фрагментов текста.

Насколько распространены словосочетания в терминологических словарях, инвентаризирующих лексический уровень текстов?

И. И. Убин приводит следующие данные о количестве словосочетаний в терминологических словарях (Научно-технический перевод, 1987):

*Таблица 2*

**Словосочетания в терминологических словарях**

| | Доля терминов в процентах длиной в: | | | |
|---|---|---|---|---|
| | 1 слово | 2 слова | 3 слова | 4 слова |
| Терминологические словари | 21 | 52 | 20 | 7 |
| Тезаурус научно-технических терминов | 26 | 40 | 24 | 10 |

Из табл. 2 видно, что доля словосочетаний в научно-технических словарях составляет 75–80 % всего состава словаря.

### 4.8.2. Автоматический анализ понятий

Среди многих проблем анализа словосочетаний выделим автоматический анализ понятий, выражаемых словосочетаниями. Этот аспект представляется нам наиболее важным по следующим причинам:

- в автоматизированных информационных системах наиболее важно определить смысловую ценность выделенной языковой единицы;
- словосочетание, определенное как некоторая единая составляющая предложения, обеспечивает переход к комплексному смысловому анализу предложения в целом, принимая при этом во внимание одновременное действие синтаксических и семантических анализаторов;
- как составляющая высказывания словосочетание прежде всего характеризуется единым смыслом;
- если мы умеем фиксировать смысл словосочетаний после выделения в тексте их границ, то словосочетание-термин также может быть идентифицировано и получает решение всех своих проблем.

Наиболее полно, на наш взгляд, проблема автоматической идентификации словосочетаний как наименований понятий получила свое освещение в работах Г. Г. Белоногова и его коллег. На материале русского языка эта проблема сформулирована им как автоматическое кодирование и декодирование наименований понятий (Белоногов и др., 1979).

Именные словосочетания — наиболее актуальный вид словосочетаний для текстов и словарей информационных систем. Именные словосочетания могут включать в свой состав следующие классы слов: существительные (С), прилагательные (П), предлоги (Р), сочинительные союзы (а) и наречия (Н). Наряду с полными буквенными кодами слов в составе именных словосочетаний встречаются также аббревиатуры, буквенно-цифровые обозначения и числа. Эти элементы обычно выступают в роли существительных и реже в роли прилагательных (например, порядковые числительные в цифровом выражении).

Количество слов в наименованиях понятий колеблется в пределах от одного до десяти-пятнадцати и в среднем равно примерно трем. Слова могут находиться в различной связи

друг с другом. Наиболее типичными видами связи являются связь согласования между существительными и определяющими их прилагательными, а также предложные и беспредложные связи между существительными.

Прилагательное, как правило, согласуется с существительным, к которому оно относится, в роде, числе и падеже. Существительное, выступающее в роли определения к другому существительному, располагается справа от последнего и может иметь форму родительного, творительного и, реже, дательного падежа. В случае предложного управления форма существительного, стоящего справа от предлога, зависит от вида последнего.

Примеры различных структур приведены на рис. 10. Здесь каждому слову наименования понятия поставлен в соответствие символ синтаксического класса. Стрелками указано направление связей между компонентами. В скобках символов существительных, не являющихся главными словами, указаны падежи, которые обозначены начальными буквами их наименований.

| No No п/п | Структурная формула | Пример |
|---|---|---|
| 1. | ПС | индикаторное устройство |
| 2. | ППС | цветное индикаторное устройство |
| ... | | |
| 4. | С → С(Р) | испытания машин |
| ... | | |
| 10. | ППС → ПС(Р) | международная автоматическая система телефонной связи |
| 11. | С → С(Р) → С(Р) | автоматизация процессов управления |
| 12. | С → С(Р) → С(Р) → С(Р) | проектирование систем обработки информации |
| ... | | |
| 20. | С → ПП | медь листовая красная и т. п. |

*Рис. 10.* **Структурные формулы словосочетаний**

Морфологический анализ является первым этапом разбора словосочетаний. Он подробно описан нами ранее (см. раздел 3). Гораздо более сложен анализ синтаксический, целью которого является установление значений связей между элементами словосочетаний. В процессе синтаксического анализа выполняются следующие операции:

- выявляется схема связей между словами;
- каждому слову словосочетания приписывается однозначная грамматическая информация, необходимая для формирования его буквенного кода при декодировании;
- структура словосочетания приводится к каноническому виду.

### 4.8.3. Анализ словосочетаний

Исходными данными для синтаксического анализа служат результаты работы алгоритма морфологического анализа. Если слова анализируются с помощью словаря словоформ, то для каждого слова наименования понятия указывается номер канонической формы слова (по словарю словоформ), набор переменной грамматической информации и постоянная грамматическая информация. Если слова анализируются с помощью словаря основ, то для каждого слова указывается номер канонической формы основы, номер флективного класса и набор переменной грамматической информации. Синтаксические связи между словами выявляются по принципу аналогии, который формулируется следующим образом: аналогичным последовательностям символов классов слов соответствуют аналогичные схемы синтаксических связей между словами. Для применения этого принципа выявляются все или наиболее часто встречающиеся в текстах последовательности символов классов слов, и им в соответствие ставятся схемы синтагматических связей. Тогда процесс синтаксического анализа сводится к распознаванию в текстах эталонных последовательностей символов классов слов. Точность анализа будет зависеть от характера принятой классификации слов, от длины эталонных последовательностей символов классов слов и от полноты представления различных синтагматических ситуаций в слова-

ре эталонов. Она будет тем большей, чем детальнее классификация слов, чем длиннее последовательности символов классов слов в эталонных описаниях синтагматических ситуаций и чем полнее словарь эталонов.

Метод аналогий целесообразно применять прежде всего для анализа текстов с ограниченными наборами синтагматических ситуаций, например для анализа именных словосочетаний.

Второй этап синтаксического анализа наименований понятий — определение однозначной грамматической информации к каждому слову. Главному слову словосочетаний — первому слева существительному — дается признак «именительный падеж», а на прилагательные переносится признак рода данного слова. Далее выделяется общая часть наборов переменной грамматической информации в группах слов, состоящих из существительных и зависящих от них прилагательных. В результате выполнения этой операции получается либо однозначная грамматическая информация, либо наборы грамматической информации, которые в дальнейшем используются для назначения информации к существительным и прилагательным.

Заключительным этапом синтаксического анализа является приведение структуры словосочетания к каноническому виду. При этом выполняются следующие операции:

- прилагательные ставятся перед теми существительными, которые они определяют;
- существительные, соединенные сочинительным союзом, располагаются по возрастанию их словарных номеров;
- группы слов, соединенные сочинительным союзом и управляемые существительными, располагаются так, чтобы управляемые слова были упорядочены по возрастанию их номеров;
- код главного слова словосочетания выносится на первое место.

Под кодированием понимается процесс замены наименований понятий на естествеином языке на некоторые формализованные смысловые коды, отражающие содержание этих понятий. Под декодированием — обратный процесс перехода от формализованных кодов к наименованиям понятий на естественном языке. Формализованный код понятия может представлять собой

его порядковый номер по списку или, в общем случае, описание его смыслового содержания на некотором формализованном языке. При этом понятие может быть описано как объект простой или сложной структуры. Если понятия кодируются их номерами, то в памяти ЭВМ целесообразно иметь два словаря: словарь слов и словарь пословных кодов словосочетаний (словарь наименований понятий). Первый словарь может быть оформлен в виде словаря словоформ или словаря основ слов. Во втором словаре каждое наименование понятия представляется сочетанием номеров слов, входящих в его состав, и номером грамматической структуры. Грамматическая структура словосочетания содержит информацию о связях между словами и информацию о формах слов, необходимую при декодировании. Различным сочетаниям номеров слов и номеров грамматических структур присваиваются порядковые номера, которые интерпретируются как номера соответствующих понятий.

Автоматическое кодирование понятий осуществляется в три этапа. Сначала отождествляются слова, входящие в наименование понятия, с элементами словаря основ. Слова заменяются их номерами по словарю и сопровождаются грамматической информацией. На втором этапе кодирования выявляется грамматическая структура наименования понятия (синтаксический анализ). Наконец, полученный в результате первых двух этапов код отождествляется с одним из элементов словаря наименований понятий и заменяется на порядковый номер этого элемента (семантический анализ). Порядковый номер понятия далее используется в качестве его кода.

### 4.8.4. Отождествление наименований понятий

Отождествление исходных и словарных наименований понятий производится в следующем порядке. Сначала сочетания номеров слов и грамматическая структура кодируемого наименования понятия ищутся по списку сочетаний номеров слов и по списку грамматических структур словаря понятий и заменяются порядковыми номерами по этим спискам. Далее по номеру понятия из словаря выбирается соответствующий ему номер грамматической структуры и сравнивается с номером,

полученным в результате поиска по списку грамматических структур. Если эти номера совпадают, то понятия тождественны друг другу. В противном случае они не тождественны.

Подобно процессу кодирования наименований понятий, их декодирование также осуществляется в три этапа. Сначала по номеру понятия из словаря выбираются соответствующие ему сочетания номеров слов и номер грамматической структуры. Затем из списка грамматических структур извлекается информация о формах слов и об их связях, а также корректируется порядок слов в словосочетании (номер главного слова ставится после номеров определяющих его прилагательных). На заключительном этапе формируются буквенные коды словоформ.

Обычно у именных словосочетаний изменяется только форма главного слова и определяющих его прилагательных. Но в некоторых случаях имеет место зависимость форм несогласованных определений и относящихся к ним прилагательных от числа главного слова (например, в словосочетаниях «директор автомобильного завода» — «директора автомобильных заводов», «начальник цеха» — «начальники цехов». Возможность такого рода преобразования не может быть обнаружена по синтаксической структуре словосочетаний. Поэтому для правильного синтеза различных форм словосочетаний необходимо в словаре наименований понятий указывать признак зависимости форм несогласованных определений от значения категории числа главного слова, а также количество существительных, на которые эта зависимость распространяется.

Наименования понятий можно представить в виде сочетаний номеров словоформ, входящих в их состав, и хранить в памяти машины два словаря: словарь пословных кодов наименований понятий и словарь словоформ. В этом случае декодирование понятий будет производиться в два этапа: сначала, с помощью первого словаря, номера понятий заменяются на их пословные коды, затем, с помощью второго словаря, пословные коды наименований понятий заменяются на их буквенные коды. Эти способы декодирования понятий довольно просты, но их применение связано с необходимостью хранения в памяти машины дополнительных словарей. Кроме того, здесь можно получить только одну форму наименований понятий.

Рассмотренные методы кодирования понятий с автоматическим отождествлением трансформационных вариантов их наименований довольно сложны в реализации и не охватывают всех видов трансформаций. Например, здесь не учитывается возможность изменения основ слов (пример: «меры защиты» — «защитные меры») и возможность изменения схем связей между словами (пример: «автоматизированная документальная поисковая система» — «автоматизированная система поиска документов» — «система автоматизированного поиска документов»). Между тем учет этих явлений весьма желателен, если в системе не накладываются ограничения на словарь входного языка. Чаще всего это бывает необходимо в документальных системах. Здесь допустимо применение упрощенных способов кодирования, при которых хотя и возможны ошибки, но зато охватывается более широкий класс трансформации словосочетаний.

Эффективным является такой способ кодирования понятий, когда слова, входящие в состав наименований, заменяются на номера смысловых эквивалентов, номер смыслового эквивалента главного слова выносится на первое место, а остальные номера смысловых эквивалентов располагаются по возрастанию их численных значений. Это дает возможность свести к одной унифицированной форме представления все трансформационные варианты словосочетаний, связанные с изменениями их синтаксической структуры, форм слов и основ слов.

Иногда возникает необходимость выборки из словаря всех понятий, подчиненных данному (всех более узких по объему понятий), или более широких, подчиняющих заданное. Эта задача может быть частично решена посредством использования синтаксической и семантической структуры именных словосочетаний. Родо-видовые отношения и отношения эквивалентности, которые нельзя выявить на основе структуры их наименований, могут быть заданы в виде специальных массивов сообщений, получивших название классификационных словарей или классификационных таблиц. В классификационном словаре каждому номеру понятия ставится в соответствие перечень номеров подчиненных и эквивалентных ему по смыслу понятий или перечень номеров по-

нятий, подчиняющих его и эквивалентных ему. Установление связей между понятиями с помощью классификационного словаря осуществляется путем выборки из него соответствующих перечней номеров понятий.

### 4.8.5. Использование словаря наименований понятий

Словарь наименований понятий и классификационный словарь можно использовать совместно в двух режимах: а) в режиме однократного обращения к словарям; б) в режиме циклического поиска. В первом случае сначала производится поиск по словарю наименований понятий, а его результаты служат исходными данными для поиска в классификационном словаре. Во втором случае после однократного обращения к словарям из общего массива результатов поиска выделяются номера терминов, полученные при поиске в классификациоином словаре и отличающиеся от номеров терминов, найденных в словаре наименований понятий. Выделенные номера терминов с помощью словаря наименований понятий заменяют их пословными кодами и обращаются повторно сначала к словарю наименований понятий, а затем к классификационному словарю. Далее среди результатов поиска по классификационному словарю снова выделяют такие номера терминов, которые не были найдены на предыдущих этапах. Эти номера заменяют на пословные коды терминов и снова обращаются к словарям и т. д. Процесс циклического поиска продолжается до тех пор, пока не перестанут находиться новые номера терминов.

Для оценки эффективности различных способов автоматического установления смысловых связей между терминами был поставлен эксперимент на ЭВМ. Для этого был использован ряд алгоритмов: алгоритм морфологического анализа, алгоритм поиска по словарю наименований понятий, алгоритм поиска по классификационному словарю понятий, алгоритм циклического поиска в классификационном словаре, алгоритм совместного циклического поиска в словаре наименований понятий и в классификационном словаре, алгоритм декодирования и оформления результатов поиска в словарях и т. д.

Исследования проводились на основе тезауруса, включавшего в свой состав около 11 700 терминов.

Были опробованы несколько способов установления смысловых связей между терминами: автономный поиск по словарю наименований понятий, построенному на основе приближенного морфологического анализа; автономный поиск по словарю наименований понятий и т. п. с использованием в разных комбинациях всех названных выше средств определения связей между терминами. Анализ результатов экспериментов установил следующее.

1. Автоматический поиск по словарю понятий дает возможность выявить только 10–14 % общего числа смысловых связей между терминами.
2. Применение для пословного кодирования терминов наряду с номерами основ слов также номеров их смысловых эквивалентов приводит к некоторому увеличению полноты установления связей (на 3,5 %).
3. Путем однократного поиска в классификационном словаре можно выявить 44–45 % смысловых связей между терминами, тогда как при циклическом поиске в этом словаре количество выявленных связей увеличивается на 20–30 % (на 20 % при поиске эквивалентных по смыслу и подчиненных понятий, на 30 % при поиске подчиняющих понятий).
4. Количество смысловых связей, выявленных при раздельном поиске в словаре наименований понятий и в классификационном словаре, меньше количества связей, выявленных при совместном однократном поиске в этих словарях (на 3–6 %).
5. Совместный циклический поиск по словарю наименований понятий и классификационному словарю обеспечивает наиболее полное выявление смысловых связей между терминами. По сравнению с совместным однократным поиском в этих словарях он позволяет выявить примерно на 40 % связей больше.

Циклический поиск в словарях во всех случаях приводит к существенному увеличению количества устанавливаемых между терминами смысловых связей. Но при этом увеличивается и уровень поискового «шума» (до 12 % при поиске эквива-

лентных по смыслу и подчиненных терминов и до 24–27 % при поиске подчиняющих терминов). Происходит это потому, что наряду с отношениями строгой эквивалентности и строгого подчинения в классификационном словаре учитываются отношения между терминами, которые не в полной мере являются отношениями эквивалентности и подчинения. При циклическом поиске неточности в связях между терминами накапливаются и уровень «шумов» возрастает.

### 4.8.6. Выводы

На этом примере конкретного подхода к построению информационно-поисковой системы с анализом и выявлением терминологических словосочетаний видны главные проблемы смыслового анализа и выявления смысла (содержания) текста посредством формального анализа словосочетаний, а также основные способы решения этих проблем. Надо заметить, что этот эксперимент, описанный в указанной выше работе Г. Г. Белоногова и его коллег, наиболее полно и исчерпывающим образом иллюстрирует методику поиска и идентификации словосочетаний терминологического характера в информационных системах. Широкий охват практического материала и промышленная постановка задачи, с высокими требованиями к результатам, дают основание верить сделанным выводам в большей степени, чем выводам из многих других теорий, не подкрепленных масштабными практическими исследованиями и разработками. Практическая задача и ее промышленное решение, результаты которого были использованы в деятельности Всероссийского института научной и технической информации (ВИНИТИ), делают описанную методику весьма актуальной в современном состоянии проблемы автоматизированного информационного поиска и содержательного анализа документации. Именно поэтому мы подробно осветили процесс обработки материала и составные части общего алгоритма решения проблемы.

## 4.9. Терминография

Терминография — наука о составлении терминологических словарей. Она является частью или разделом лексикографии, общей науки о словарях, и в то же время она тесно связана с терминоведением.

Терминография — терминологическая лексикография — является следствием и одновременно необходимым условием научно-технического прогресса. В терминологических словарях фиксируются элементы научного знания, без которого научные и технические исследования и разработки были бы невозможны.

Согласно исследованию, проведенному Всесоюзным центром переводов научно-технической литературы и документации СССР в 1986 году, в мире каждый день публикуется один специальный словарь. Терминологические словари разнообразны: есть одноязычные словари, задачи которых также весьма многообразны, двух- и более язычные словари, которые предназначены для научно-тенического перевода, являющегося важнейшей компонентой научно-технического прогресса. Количество терминологических словарей и их виды, выпускаемые в какой-либо стране, достаточно полно характеризуют научные приоритеты этой страны.

Весьма важен тот факт, что лексическое богатство каждого естественного языка из века в век увеличивается. Первый словарь английского языка, опубликованный как результат научного исследования языка в 1604 году, содержал 3000 слов. Словарь Джонсона 1750 года имел уже 43 500 слов. В 1973 году Краткий Оксфордский словарь включает уже 163 000 слов. Во Франции универсальный словарь Фуретье 1690 года включал 50 000 слов. Словарь Французской академии 1964 года содержит 25 000 слов только высокочастотного характера. Энциклопедия Гран ЛяРюс в десяти томах содержит каждое слово французского языка, а всего 450 000 слов. Второе издание Малой Советской Энциклопедии вышло в 1933–1940 годах и содержало 31 000 словарных статей. Третье издание Малой Советской Энциклопедии, рассчитанное на широкие круги читателей — рабочих, колхозников, городскую и сельскую интел-

лигенцию, учащуюся молодежь, вышло в 1958–59 годах и включало уже около 50 000 словарных статей. В ней освещены основные понятия и термины, встречающиеся в современной научной, художественной и публицистической литературе. При этом она не заменяет собой толковый словарь русского языка. «Толковый словарь русского языка» С. И. Ожегова и Н. Ю. Шведовой издания «АЗЪ» 1992 года содержит 72 500 слов и 7500 фразеологических единиц. Таким образом, русский язык, который до недавнего времени находился на втором месте в мире по числу публикуемых на нем изданий, также регистрирует значительный прирост в количестве слов. При этом русский язык широко использует словообразовательные потенции языка (Марчук, 2004).

Что касается терминологии, то она характеризуется количественным ростом, описываемым экспоненциальной кривой. В начале XX века вся научная терминология немецкого языка, наиболее тщательно регистрировавшего научные и технические термины, насчитывала около 3,5 млн. терминов. В восьмидесятые годы в области только одной электротехники в немецком языке более 4 млн. терминов (Better Translation, 1983).

Представляет интерес распределение терминологических словарей в мире. Так, США занимают первое место по выпуску одноязычных словарей в течение 1950–1979 годов (23 % общего мирового объема) и на четвертом месте по выпуску переводных словарей. По сравнению с другими странами, Германия характеризуется активным участием в выпуске различных типов словарей и определенной сбалансированностью всех типов словарей (одноязычные словари 64 %, двуязычные 23 %, многоязычные 13 %). Участие Франции в производстве терминологических словарей увеличивается с начала 1970-х годов, и она занимает третье место. Во Франции словари посвящены самым различным предметным областям: археологии, психологии, геологии, электронике и т. д. Большое значение придается медицине.

С точки зрения научного и технического прогресса весьма интересен пример Японии. До 1950 года терминологическая активность Японии была незначительной. Однако после

1950 года отмечается резкий подъем, бум словарного дела, который предшествовал научному и техническому развитию страны. В течение всего исследованного ВЦП периода (1950–1979 годы) выпускается множество словарей по строительству, химии, металлургии, физике, автомобильной промышленности и т. д. Эта деятельность способствовала распространению английского языка среди специалистов Японии (большинство терминологических словарей было с участием английского языка) и увеличивала знания в науке и технике. Этим подготавливался известный рывок Японии в промышленности и науке в начале восьмидесятых годов XX века.

В Китае отмечается охват терминологическими словарями различных областей знания. В настоящее время активно составляются словари на машинных носителях по различным областям машиностроения, робототехники, нефтехимии, биологии и другим наукам.

Индия представляет собой типичный пример развивающейся страны. Преобладающий тип словарей — двуязычные словари для перевода. Главную роль играют фундаментальные науки. В начале 1960-х годов появляются словари не только на хинди, но и на бенгали.

Новые науки характеризуются выпуском возрастающего числа терминологических словарей. Это необходимо для терминологической работы и для развития новых областей знания, а также новых отраслей промышленности и прогресса. В России в настоящее время также отмечен бум в выпуске терминологических словарей. Выходят новые словари по маркетингу, страховому делу, юридической и судебной практике, нефтехимии, программированию и вычислительной технике и по многим другим областям деятельности, как научно-технической, так и социальной. Возобновляют работу организации, занимающиеся нормированием и упорядочением терминологической деятельности.

Тематика терминологических словарей чрезвычайно важна для наукометрии, так как позволяет определить области науки и техники, в которых идет интенсивное развитие.

### 4.9.1. Методы терминографии

Терминография представляет собой специальный вид деятельности, направленной на регистрацию отношений «термин-концепт» с помощью всех терминологических данных. Всякие другие данные, ассоциированные с терминологическими данными, также могут так или иначе отражаться в терминологических словарях и применяться в терминографии.

Обычно говорят о двух типах терминографии: терминографии дескриптивной, задачей которой является найти наиболее точные описания терминов, и терминографии прескриптивной, задачей которой является предписать, какой термин следует использовать. Существует также и представление о том, что в научно-технической коммуникации невозможно предписывать использование терминов, поэтому задачей терминографии является распространение и пропаганда терминов и соответствующих систем терминов — терминологий — с тем, чтобы ученые сами могли выбирать, какими терминами пользоваться в каждом конкретном случае, когда возникает такая необходимость.

Методы терминографии включают:

- выбор терминографических данных;
- размещение лексических единиц в специализированных словарях, документационных тезаурусах или вокабулярах;
- организацию и структуру специальных терминологических словарей;
- размещение частей терминологического словаря относительно друг друга.

Различные задачи, такие, как передача знаний и технологий, документация, информация и научно-технический перевод, требуют специальных средств для своей реализации. Упорядоченное собрание лексикографических данных может выступать в качестве вокабуляра, словаря, картотеки, банка данных, документационного тезауруса или быть представленным в другом виде.

Базовой единицей терминографической коллекции данных является терминографическая единица, которая содержит терминографическую информацию в форме, наиболее удобной для данной задачи.

### 4.9.2. Основные требования к специальным словарям

Оценка эффективности терминологических словарей связана с требованиями, предъявляемыми к этим словарям. Обычно такие требования отражают специфические нужды пользователей. Однако можно выделить и общие требования к специальным словарям терминологий. Это следующие требования:

- адекватное описание лексики выбранной области науки и техники;
- наличие всей необходимой информации;
- отсутствие избыточной и ненужной информации;
- унификация композиции и систем индексации подобных словарей с тем, чтобы облегчить переход от одного словаря нужного типа к другому. В некоторых случаях создаются системы словарей или несколько словарей сходного типа, имеющие одну цель и одно терминологическое выполнение. Например, французско-англо-русско-кхмерский географический словарь, французско-англо-русско-кхмерский лингвистический словарь, французско-англо-русско-кхмерский исторический словарь. Все эти словари создавались в рамках единого словарно-терминологического проекта Министерства образования Камбоджи в 1991 году.

Словарь для перевода прежде всего адресован профессиональным переводчикам. Для перевода самое важное — наиболее быстро найти точный переводной эквивалент искомому слову в выходном языке. Эти два параметра — скорость обнаружения перевода и его точность — лежат в основе всех требований к словарям такого типа.

Толковый словарь может быть представлен специализированным словарем узкой предметной области. Такой словарь предназначен для экспертов в конкретной научно-технической области. Это словарь справочного типа, его пользователи обычно ищут в таком словаре точные определения терминов, чтобы понять их смысл и семантические аспекты их использования. Примером словарей такого типа может служить строительный словарь (Строительный словарь, 1985).

Учебный словарь рассматривается как средство овладения лексикой иностранного языка. Главной целью такого словаря яв-

ляется помочь изучающему понять семантические отношения между различными словами и терминами словаря. Именно поэтому в таких словарях большую роль играют объединения слов в семейства. Учебный словарь всегда является словарем нормативным, поскольку изучение использования слов основывается на презумпции некоторого семантического порядка и семантических связей между словами. Выбор слов в таком словаре должен прежде всего отражать требования к обучению языку.

Все эти общие требования уточняются применительно к конкретным словарям, составляемым для решения специфических задач научно-технической и учебной коммуникации, где используется терминология.

### 4.9.3. Многоязычная лексикография и терминография

Научно-технический и социальный прогресс в современном мире характеризуется активизацией роли развивающихся стран и появлением новых национальных языков в мировом общении. Соответственно возрастает роль многоязычных словарей, которые занимают особое место как в лексикографии и терминографии, так и в соотношении культур.

Вопрос о соотношении культур языков, вступающих между собой в контакт, весьма важен для перевода и составления словарей. В настоящее время вопросам культуры уделяется все больше внимания в аспекте отношения языка и культуры (см., например: Рождественский, 2003; Гуревич 1996 и пр.). Набор значений слов словаря отражает соответствие понятий разных языков, в каждом из которых свой культурный уровень и свой набор мыслительных сущностей, долженствующих быть выраженными словами. В двуязычной ситуации действуют два культурных уровня. В многоязычной — три и более, и взаимодействия между ними достаточно сложны, во всяком случае они значительно влияют на проблематику составления терминологических словарей.

До последнего времени зоной распространения многоязычных словарей обычно считалась многоязычная Европа и некоторые многоязычные страны, например Канада. Немало много-

язычных словарей составлено по политехнической тематике, создаются также многоязычные банки терминологических данных.

Одно- и двуязычная лексикография по общему вниманию к ней превосходит многоязычную терминографию и лексикографию. В последней много неясного: так, известно, что при увеличении числа языков уменьшается возможность давать информацию пояснительного характера в каждой словарной статье. Перевод и другая работа с иноязычными текстами обычно включают два языка и редко — большее их число. В многоязычной лексикографии обычно остро стоит вопрос о многозначности слов. Если языки, охватываемые словарем, принадлежат к разным языковым семьям, то возникает еще немало других дополнительных проблем, усложняемых разной письменностью, трудностью размещения лексических единиц, составления индексов и пр. Вследствие этого многоязычная лексикография и терминография представляют собой довольно мало разработанную часть общей и переводной лексикографии и терминографии.

Актуальность многоязычной терминографии резко возрастает с появлением автоматизированных словарей и банков терминологических данных. Эти устройства решают многие проблемы лексикографии и терминографии: так, отпадает необходимость экономить место, поскольку память на машинных носителях достаточно велика; появляются новые способы удобного соотнесения эквивалентов на разных языках и т. п. Тем не менее, основные содержательные проблемы многоязычной лексикографии не снимаются скоростью действия и большим объемом памяти компьютеров.

Многоязычные словари чаще всего предназначены для перевода. В условиях многоязычия развивающихся стран эти словари также играют важную роль в становлении национальной терминологии. Цели и задачи многоязычного словаря являются определяющими во многих отношениях. Они служат основанием для выбора объема словаря, уточнения тематической области, состава источников, выбора структуры словаря и методики его составления.

Составители многоязычных словарей исходят из предположения о том, что человечество имеет одни и те же корни,

природа человеческих органов речи и интеллекта человека одинакова для всех народов, поэтому в языках человека много общего как в лексике, так и в грамматике (Drouin, 1866). Многоязычные словари составлялись уже довольно давно, однако терминологические многоязычные словари — явление сравнительно недавнее. Если задачей универсального многоязычного словаря общеупотребительной лексики является сравнение языков на лексическом уровне и создание некоторого общего понятийного уровня для разных языков, исходя из принципа единства «корней» человеческого мышления, то многоязычный терминологический словарь имеет целью обозначить во многих языках единые границы терминологического поля как определенных предметных областей, так и конкретных более или менее общих понятий или концептов в терминологии.

Словари-источники для составления многоязычного словаря можно разделить на следующие категории:

- одноязычные толковые и энциклопедические словари;
- двуязычные терминологические словари как узкоспециальные, так и политехнические;
- многоязычные терминологические словари общего характера (например, политехнические).

Роли этих словарей разные. Энциклопедический или толковый словарь, содержащий основной список терминов с их толкованиями на одном каком-либо языке, служит обычно основой словника, предметным ориентиром, с помощью которого определяются границы предметного поля и сам состав словника. С другой стороны, энциклопедии и толковые словари необходимы для определения содержания понятия в случае, когда возникают неясности или когда в разных языках содержание понятия различно. При составлении Институтом научных исследований Камбоджи словаря географических терминов на четырех языках в качестве исходных служили два современных энциклопедических словаря — Dictionnaire de la Géographie, 1984 и Penguin Dictionary, 1987. Двуязычные терминологические словари дают возможность выбрать конкретный переводной эквивалент. Многоязычные словари служат той же цели, и, кроме того, с их помощью можно проверить выбранный эквивалент Весьма полезным при этом оказывается раз-

ноязычие словарей, поскольку возникает возможность лучшей перекрестной проверки.

Составителю многоязычного словаря приходится разрешать конфликты между разными словарями в подаче переводных эквивалентов. Так, при составлении географического словаря в процессе поиска переводных эквивалентов было обнаружено, что французско-русский политехнический словарь для французского слова résurgence дает перевод «выход подземных вод на поверхность». Французско-русский геологический словарь переводит этот термин как «область разгрузки подземной реки». Принимая во внимание то обстоятельство, что составляемый географический словарь имеет широкую специализацию, т. е. не является собственно геологическим, было решено воспользоваться эквивалентом политехнического словаря.

Терминологическая номинация — целенаправленный творческий процесс, обусловленный взаимодействием внешних и внутренних языковых факторов. Основой создания любой терминологии является использование национального языкового фонда. Формирование новых терминов на базе национальных языковых традиций осуществляется путем семантического словообразования (использование общеупотребительных слов в функции терминов), а также с помощью корневых и аффиксальных морфем по словообразовательным моделям, типичным для соответствующего языка. Лексико-семантическое образование русской терминологии подробно исследовано в капитальной работе В. Н. Прохоровой (Прохорова, 1996). Если термины разных языков обозначают один и тот же предмет или явление, но имеют различные признаки номинации, соотносимые с разными сторонами именуемого предмета или явления, то такие термины как бы дополняют друг друга с точки зрения того, какие стороны объективной реальности в них фиксируются. Этот же факт, т. е. отражение различных картин мира в национальной языковой номинации, служит причиной многозначности терминов в многоязычных терминологических словарях.

Проблема многозначности терминов в многоязычных терминологических словарях была изучена Н. Н. Занегиной на материале семиязычного словаря терминов рекламного дела. В трехъязычном выпуске словаря (английский, немецкий, ни-

дерландский языки) она отобрала 10 многозначных слов, имеющих от 2 до 5 значений. Очень убедительно по всем отобранным словам-терминам показан разброс значений по языкам, точнее, различных оттенков значений. Так, немецкий язык наиболее абстрактен в определениях, нидерландский — более конкретен, английский занимает промежуточное положение. Практическим следствием такой ситуации является трудность выбора правильного русского перевода для того или иного многозначного термина (Занегина, 2001). В теоретическом же плане проблема многозначности многоязычных терминов как весьма актуальная для современной информационной обстановки подлежит дальнейшему исследованию.

### 4.9.4. Какие аспекты терминографии актуальны для информатики

Вышесказанное относительно терминологии, терминоведения и терминографии актуально прежде всего для компьютерной лингвистики. Однако следует подчеркнуть тот факт, что для информатики как специализированной науки об извлечении информации из текстов на естественных языках прежде всего важна форма, так как именно от формы лингвистических данных, от плана выражения, отправляется любая информатическая процедура.

Форма эта — прежде всего форма слова, словоформа, т. е. та форма, в которой данное слово встретилось в тексте. Это касается также и тех языков, для которых слово, сочетание слов, морфема могут встречаться в тексте с равными вероятностями (китайский язык, например).

Неоднозначность формы может быть разного рода, и в вычислительной, компьютерной лингвистике ряд традиционных лингвистических понятий, такие, как словоизменение и словообразование, лексическая омонимия и лексическая многозначность, теряют свое различие. Эта неоднозначность разрешается различными средствами, которые зависят от целей и задач, в свою очередь определяющих состав и структуру лингвистического моделирования, природу тех моделей, которые выбираются для осуществления на компьютере соответствующих лингвистических функций человеческого мышления.

# *Глава 5*
# Моделирование в компьютерной лингвистике

В данном разделе мы рассмотрим проблемы, связанные с изучением и использованием в целях компьютерной (прикладной, вычислительной) лингвистики языковых единиц, больших, чем морфема, слово, словосочетание. Речь пойдет о предложении, высказывании, синтагме и пр., то есть о лингвистических единицах, выражающих то, что можно определить как некоторую законченную мысль. Если все предыдущие единицы относились в общем случае как бы к имени, то есть являлись результатом номинации, простого (или сложного) наименования объектов высказывания и мысли, то теперь мы переходим к единицам, содержащим предикативность, то есть есть имя, о котором что-то сказано, констатировано наличие признака, действия, состояния и пр. В языке подобная мысль выражается высказыванием, более или менее законченным. Частично вопрос о содержании высказывания был рассмотрен в разделах о синтаксическом и семантическом анализе. Однако незначительные результаты в формализации синтаксического и семантического анализа в прикладных целях объясняются тем, что эти уровни тесно связаны с мышлением человека. Поэтому в данном разделе речь пойдет о соотношении языка и мысли, об интеллекте, о моделировании интеллекта и создании систем «искусственного интеллекта». Мы не будем затрагивать философских аспектов этой проблемы. Нас интересует прежде всего прикладное, лингвистическое содержание проблемы, возможности использования компьютера для решения лингвистических задач, так или иначе связанных с мышлением.

## 5.1. Моделирование языковых сущностей и человеческого мышления

### 5.1.1. Связь языка с мышлением

Связь языка с мышлением — сложнейший философский вопрос, над которым работают ученые с древних времен до наших дней. Мы исходим из предположения, что такая связь есть, она является решающей и принимает такие формы, которые можно использовать для более или менее успешного формального описания высказывания, предложений и текста. «В том, что язык является средством формирования мысли, нет никакого сомнения. После В. Гумбольдта эта идея для всех стала привычной. Однако при конкретном исследовании языка и мысли большинство авторов обнаруживают между ними такое глубокое различие, что невольно начинают рассматривать язык и мысль в виде двух параллельных взаимосвязанных потоков. И какие бы при этом не произносились слова о неразрывном существовании языка и мысли, представление об оголенном существовании мысли становится неизбежным, мысль оказывается существующей хотя и в какой-то связи с языком, но лишь рядом с ним и вне его». «Учение о функциях языка, которое призвано раскрыть непосредственную связь языка с мыслью, является малоразработанной областью в науке о языке» (Наседкин, 1967, с. 65). Далее автор говорит, что наиболее разработанной является функция наименования, а семасиологическая разработана крайне мало. Это крайняя точка зрения на связь языка и мышления.

Другой точки зрения придерживается А. С. Чикобава. «Для нас язык — это система знаков, функционирующих в качестве средства общения и инструмента мысли» (Чикобава, 1967, с. 18). В языке этот автор выделяет две функции: интериндивидуальную — быть средством общения, коммуникации, и интраиндивидуальную — быть выражением мысли. Ведущая роль принадлежит первой: язык, переставший быть средством общения, считается мертвым, он не может сохранять интраиндивидуальную функцию. Никому не удается долгое время мыслить на том языке, который нет возможности использовать в качестве средства общения.

Определяющая роль коммуникативной функции демонстрируется тем самым ярко: интериндивидуальная функция служит опорой интраиндивидуальной. Такова еще одна точка зрения на связь языка и мышления.

А. Т. Кривоносов, рассматривая проблему частей речи с теоретической точки зрения, включает ее в структуру «действительность-мышление-сознание-язык». Понятие «мышление» в языкознании, логике, психологии, философии настолько неопределенно и так запутанно, говорит он, что под «мышлением» некоторые исследователи иногда понимают чуть ли не противоположное. Прежде всего потому, что само содержание термина «мышление» расплывчато и не имеет четких границ (Кривоносов, 2001, с. 144).

По степени, глубине, силе абстракции и обобщения необходимо различать, по А. Т. Кривоносову, два уровня мышления или две ступени познания: 1) чувственное (непосредственное, наглядное, эмпирическое, техническое мышление, осуществляющееся на первой ступени абстракции) в формах ощущений, восприятий, представлений, которые не требуют языковой опоры (первая сигнальная система, по И. П. Павлову), и 2) абстрактное, логическое, теоретическое мышление, осуществляющееся на второй ступени абстракции (соответственно, по И. П. Павлову, вторая сигнальная система). Это мышление реализуется в логических формах понятий, суждений, умозаключений, обязательно выраженных в формах естественного языка — в словах (понятиях) и предложениях (суждениях, умозаключениях). Логическое мышление также может быть выражено и вне форм языка, это происходит тогда, когда мы думаем, рассуждаем про себя, вспоминаем о чем-нибудь, не прибегая к формам языка. Именно безъязыковое (авербальное) мышление и является, в сущности, типичной формой человеческого мышления.

Две ступени познания — чувственное созерцание, осуществляющееся посредством органов чувств, и абстрактное, логическое мышление, обобщающее данные чувственного познания, образуют единый процесс мышления, в результате чего происходит постоянное превращение элементов или континуума чувственного мышления в абстрактное мышление и аб-

страктного мышления в чувственное мышление. Далее различаются два уровня движения логического мышления: «процесс мышления» (динамика мышления) и «сознание» (статика мышления). В чувственном и логическом мышлении мы обнаруживаем сложное взаимодействие между реальной действительностью и мозгом, а в логическом мышлении — также и с языковыми знаками

А. Т. Кривоносов далее в этой работе критикует «когнитивную лингвистику» за то, что она только запутывает старые проблемы языкознания. Для того чтобы разобраться в основных понятиях мышления, рассмотрим общепринятые точки зрения на такие предметы, как искусственный интеллект и процесс мышления у человека.

### 5.1.2. Элементы системы искусственного интеллекта

В процессе умственной целенаправленной деятельности человек формирует подлежащие решению задачи, отыскивает правила, учитывает разного рода обстоятельства, наконец, вырабатывает решение, отбрасывая ненужную информацию. Все эти элементы мыслительной деятельности должны как-то закладываться в программу компьютера, если мы хотим, чтобы он обладал «мышлением». Модель функционирования системы искусственного интеллекта (ИИ) может быть представлена так, как показано на рис. 11.

Приведенная на этом рисунке структура компьютерной модели ИИ основана на раздельном функционировании различных компонентов системы (Ефимов, Фролов, 1991). Этот самый важный фактор и делает программу ИИ гораздо более эффективной по сравнению с другими, обычными компьютерными программами. Эта концепция структурной автономности очень важна и с другой точки зрения. Поскольку элементы программы отождествляются с «составными частями» человеческого мозга, можно заложить в программу ИИ сам способ мышления высококвалифицированного специалиста по любой отрасли знания. Если можно определить, что «делается» в мозге на каждой стадии этого процесса, мы можем

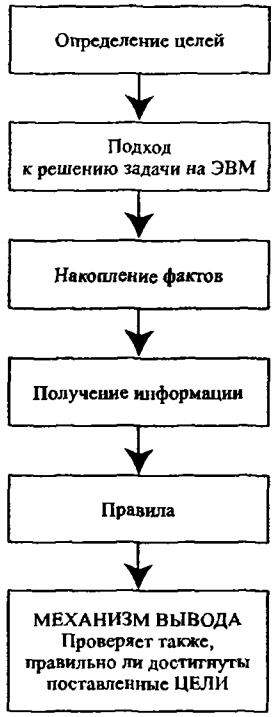


*Рис. 11.* **Основные элементы системы искусственного интеллекта**

легко подыскать эквивалентный участок программы, который соответствует такому же участку в человеческом мышлении. Поскольку достижение цели является задачей любой системы ИИ, первым шагом в ее создании должно быть точное определение этих целей. Необходимо знать, какого рода проблему вы хотите решать, и суметь ее описать, пользуясь конкретными обозначениями и терминами, прежде чем дать соответствующее задание компьютеру.

### 5.1.3. Как мыслит человек

Под системой с названием «искусственный интеллект» в практическом плане обычно понимают компьютерную про-

грамму, способную «думать» и решать так называемые «творческие» задачи.

Для реализации (искусственного интеллекта — ИИ) надо прежде всего изучить, как мыслят люди, когда им надо принять какое-либо решение или решить ответственную проблему. Следует выделить в мыслительном процессе основные стадии, что и позволит в дальнейшем разработать компьютерную программу, способную решать задачи с использованием тех же стадий мыслительного процесса. ИИ, таким образом, обеспечивает, по крайней мере в принципе, простой структурный подход к разработке весьма сложных программ, позволяющих решать творческие задачи весьма разного уровня.

Стандартная компьютерная программа способна обеспечить ответы только на те вопросы, для которых она специально создана. Если требуется изменить стандартную программу для пополнения ее новой информацией, то приходится ее тщательно просмотреть, чтобы найти место для новой информации. Это не только отнимает время, но может и нанести вред программе, которая может теперь давать ошибки из-за влияния на другие части программы. Искусственный интеллект, как и предполагает название, действительно делает компьютер способным думать. ИИ имитирует основной человеческий процесс обучения, при котором происходит прием информации и ее усвоение для дальнейшего использования. Человеческий мозг способен воспринимать все новые и новые знания без изменения процессов жизнедеятельности и без функциональных нарушений различных отделов головного мозга. Программа ИИ действует почти таким же образом.

Каждый структурный блок программы напоминает фрагмент информации, заключенный в человеческом мозге. Если эта информация неточна или сомнительна, мозг может автоматически приспособить мышление для восприятия нового ряда фактов. При этом требуется восстановить в памяти не всю информацию, а только те ее части, которые требуют корректировки или замены новой информацией.

Разумеется, стандартная программа может выполнить в принципе почти все, что и программа ИИ, но составить ее нельзя так же легко и быстро, говорят Ефимов и Фролов.

В обоих типах программ есть блоки, которые независимы в смысле выполнения разработанных пользователем функций. Но программа ИИ обладает одной отличительной чертой, которая приравнивается к черте человеческого интеллекта. Любое звено в программе ИИ может быть изменено и модифицировано без вредного воздействия на структуру всей программы. Такая гибкость обеспечивает большую эффективность программирования и способность понимания пользователем всех шагов ЭВМ. Иными словами, компьютер обретает «интеллект».

Так как ИИ — сфера изучения наук, базирующихся на специфике человеческих мыслительных процессов, то очень важно углубленное исследование этих процессов. Пока никто точно не знает, как работает мозг. Человеческий интеллект — это сложная социально-биологическая функция, которую ученые только начинают постигать.

## 5.2. Искусственный интеллект

Литература по искусственному интеллекту в настоящее время достаточно обширна (Поспелов, 1983; Шабанов — Кушнаренко, 1984; Потапова, 1989; Потапова, 2002; Зубова, 2001; Рябцева, 2002 и др.). В данном параграфе мы рассмотрим некоторые вопросы моделирования мышления в целом, прежде чем в дальнейшем перейти к моделированию лингвистических способностей человека в части синтаксиса и семантики.

### 5.2.1. Модель механизма мышления

Любого человека легко поставить в ситуацию, когда он не сможет адекватным образом отреагировать на происходящее (выполнить нужное действие, правильно ответить на вопрос и пр.). Умение правильно и своевременно реагировать в самых различных ситуациях у человека и животных приобретается лишь благодаря способности к обучению, без которой было бы невозможно развитие человека (Вайнцвайг, Полякова, 1987). Эта способность, в свою очередь, тесно связана с требованием работы в реальном времени, поскольку в результате обучения

появляется возможность не только решать новые, ранее недоступные задачи, но и быстро решать те задачи, которые раньше решались медленно.

Оптимальная с точки зрения выживаемости организма конструкция механизма мышления должна удовлетворять следующим условиям:

- работать в реальном времени;
- обладать способностью к обучению;
- максимально полно использовать поступающую извне информацию как на этапе принятия решений, так и на этапе обучения;
- иметь полную память о прошлых событиях и обладать способностью непрерывного обобщения поступающей информации.

С помощью одного процессора такие условия трудно выполнить, поскольку для их выполнения требуется практически мгновенная обработка громадных массивов данных. Требуется использовать параллельные вычисления. Проект обучаемой системы управления поведением, изложенный в цитируемой работе, предусматривает использование параллельно работающей ассоциативной памяти — процессора. На рис. 12 изображена модель механизма мышления.

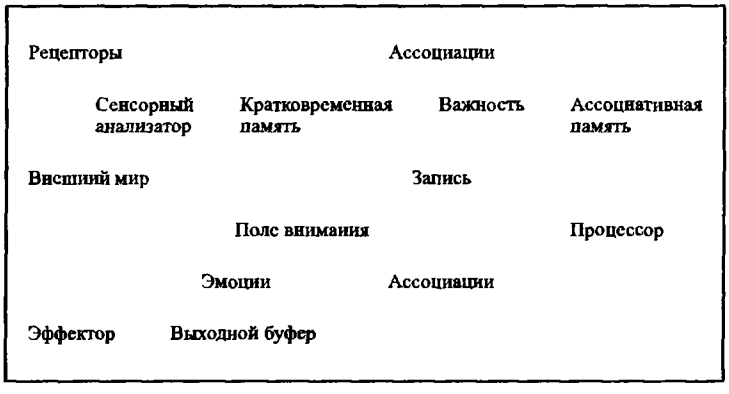


*Рис. 12.* **Модель механизма мышления**

Блоки системы выполняют следующие функции.

Информацию о внешнем мире и о своих внутренних состояниях система получает с помощью рецепторов, а посредством эффекторов она выполняет те или иные внешние действия. Поведение системы в общем случае управляется двумя функциональными блоками: сенсорным анализатором и ассоциативной памятью-процессором. Сенсорный анализатор — это устройство, осуществляющее обработку входной (сенсорной) информации, основанную на жестких, необучаемых механизмах. Сюда входят, в частности, механизм зрительной константности, механизм бинокулярного стереосинтеза и восстановления формы предметов по движению, механизм слуховой ориентации и пр. Сенсорный анализатор осуществляет согласование информации, получаемой от различных органов чувств, и обеспечивает константное восприятие важных свойств и пространственно-временных характеристик предметов. Сенсорный анализатор в результате своей работы реализует следующие функции:

- формирует первичные оценки качества ситуаций — первичные эмоции (меру удовольствия или неудовольствия и пр.);
- при выполнении соответствующих условий формирует команды безусловно рефлекторных (инстинктивных) действий;
- при выполнении определенных условий формирует те или иные цели (построить гнездо, добыть пищу и пр.);
- строит описание текущего состояния организма в терминах его ощущений, производимых действий, поставленных целей и испытываемых эмоций, которое (описание) через поле внимания может быть занесено в основную (долговременную) память и использовано для обучения и формирования дальнейшего поведения;
- выдает команды управления «внутренними эффекторами», т. е. управляемыми параметрами самого сенсорного анализатора, такими, например, как размеры и форма рецептивных полей, типы детекторов, размер эффективного поля зрения и пр.;

- для каждого из результатов вычисляет значение их важности.

Выходной буфер производит по максимуму важности выбор команды, поступающей либо от сенсорного анализатора, либо от основной (ассоциативной) памяти и разворачивает ее в программу действий, выполняемых эффекторами.

Блок эмоций суммирует оценки ситуации, получаемые от сенсорного анализатора (первичные эмоции) и оценки, получаемые из основной памяти и зависящие от ее состояния (вторичные эмоции). Суммарная оценка служит мерой качества текущего состояния организма. Она постоянно поступает в поле внимания и запоминается вместе с другой заносимой в память в данный момент информацией, в результате чего все события, сохраняющиеся в памяти, оказываются эмоционально окрашенными.

Кратковременная память является местом, где хранится информация, поступающая от сенсорного анализатора, пока в основной памяти осуществляется коррекция ее важности, и откуда эта информация может потом попасть в поле внимания, если ее важность окажется достаточно большой.

Поле внимания представляет собой регистр (или канал) основной памяти, на входе которого в каждый момент времени по максимуму важности производится выбор информации, поступающей либо от сенсорного анализатора, либо из памяти. Объект, попавший в поле внимания, всегда запоминается, причем запоминаются все те и только те объекты, которые проходят через поле внимания.

Основная память, или ассоциативная память-процессор, является как местом постоянного хранения информации, поступающей через поле внимания, так и процессором, осуществляющим обработку этой информации, в основе которой лежит операция ассоциации, т. е. сопоставление описания текущей ситуации со всем содержимым памяти. Память имеет два ассоциативных входа, на один из которых информация поступает от сенсорного анализатора, на другой — из памяти.

В процессе ассоциации осуществляются: формирование понятий и законов, постановка задач, решение задач и формирование вторичных эмоций. Функциями ассоциативной памя-

ти-процессора в основном и определяется весь мыслительный процесс.

Слово «осознать» можно рассматривать как синоним выражения «обратить внимание на ассоциируемый объект», а слово «сознание» — как синоним «память об ассоциации». В частности, если вспоминается (т. е. вновь поступает в поле внимания и заносится в память с модальностью «из памяти») только что виденный объект, говорят об осознании факта видения этого объекта; если образ конкретной собаки вызывает в памяти ассоциацию с ранее сформированным обобщенным понятием «собака», говорят об осознании того, что в данный момент видится собака. Лишь при наличии осознания, которое позволяет производить перегруппировку в описании естественного хода событий и соединять вместе разбросанные по памяти элементы, появляется возможность сопоставлять и находить общее в различных процедурах поиска, путях рассуждения, решении задач и пр. Осознание возможно лишь при наличии канала, по которому информация из памяти может поступать в поле внимания и таким образом может быть еще раз записана на новом месте памяти. Именно это отличает сознательную систему от условно-рефлекторной.

### 5.2.2. Ассоциативное построение понятий

В памяти имеется информация двух типов: 1) основные данные, к которым относятся описания ощущений, эмоций, выполняемых действий, законов, задач и путей их решения, словом все, что проходило когда-то через поле внимания, сохраняется в неизменном виде, а потому всегда доступно сознанию; 2) соответствующие основным данным вспомогательные параметры, такие как величины ассоциации, важности, надежности предсказания и пр., постоянно перевычисляемые в процессе ассоциаций и используемые в процессе обучения, постановки и решения задач, возникновения вторичных эмоций, переключения внимания и пр.

Ассоциация изначально ориентирована на формирование и работу с выражениями некоторого языка, на котором описываются понятия (предикаты от ситуации), а также законы и за-

дачи. Такой язык должен удовлетворять следующим требованиям. Ассоциация должна быть в равной мере применима как к исходным описаниям, так и к сформированным вариациям этого языка. В частности, должна реализовываться возможность выяснения применимости к конкретному описанию тех или иных обобщений. Кроме того, выражения языка должны достаточно просто формироваться в процессе ассоциации.

От выбора конкретного языка и критериев формирования его выражений в значительной мере зависит эффективность обучения и работы системы. Предварительно в рассматриваемой модели в качестве такого языка рассматривается расширение языка регулярных выражений.

Понятия и законы внутреннего языка в системе предполагается формировать рекурсивно, когда сложные понятия строятся на основе ранее построенных простых понятий. Идея состоит в том, что понятия ограниченной сложности могут строиться в процессе последовательных ассоциаций непосредственно в тех же местах памяти, где хранятся исходные описания.

Ассоциативное построение понятий предполагается уложить в следующую простую схему.

Каждый элемент содержимого памяти, сопоставляясь с элементом входной последовательности, может интерпретироваться фиксированным числом способов — быть отсылкой к какому-либо уже построенному понятию; быть началом, продолжением или концом конкретного слова; началом, продолжением или концом переменного слова, которое может быть произвольным или иметь фиксированную длину; быть связанным с соседним словом упорядоченной или неупорядоченной конкатенацией и т. д. Каждому из этих способов интерпретации соответствует вес, который служит коэффициентом при выборе одного из возможных путей ассоциации. Набор таких весов называют ассоциатором данного элемента памяти.

В результате ассоциации в каждом из ассоциаторов, принадлежащих второму этажу памяти, пропорционально энтропии распределения его весов, величине ассоциации и надежности предсказания последующих элементов памяти происходит относительное увеличение весов тех способов интерпретации,

которые принимали участие в ассоциации, т. е. принадлежат ее следу.

Последовательность значений ассоциаторов на некотором отрезке содержимого памяти представляет собой как бы текущее значение обобщенного описания ситуации — своеобразное нечеткое понятие, эквивалентное некоторому множеству предикатов от ситуации с заданным на нем распределением весов. В процессе последующих ассоциаций на отдельных участках памяти будет происходить постепенное контрастирование распределения весов различных способов интерпретации, приводящее к выделению среди них наиболее предпочтительных.

В сформированном виде, когда состояния ассоциаторов определяют лишь единственно возможную последовательность типов интерпретации, обобщенное описание ситуации соответствует тому, что обычно называют понятием, и формально эквивалентно одиночному предикату от ситуации.

Одновременно с построением понятий происходит формирование законов. Они строятся в виде утверждений, имеющих вид причинно-следственной связи.

### 5.2.3. Основные принципы работы системы и организации ее поведения

Существуют три уровня организации поведения:

- безусловно-рефлекторный (инстинктивный) уровень, характеризуемый тем, что отсутствует использование каких-либо сведений из основной памяти;
- условно-рефлекторый (бессознательный) уровень, характеризуемый тем, что при организации поведения хотя и используются сведения из основной памяти, но никакие промежуточные результаты процесса этой организации не осознаются, т. е. не попадают в поле зрения и не запоминаются;
- сознательный уровень, основанный, как правило, на многократном (циклическом) использовании памяти, когда результаты работы каждого шага цикла, попадая в поле внимания и вызывая в памяти дальнейшие ассоциации, запоминаются.

Подробнее о сознательном поведении авторы говорят следующее. Подобно тому, как в обычной вычислительной машине не любая функция может быть реализована без использования дополнительных полей памяти, не любая поведенческая задача может быть решена на условно-рефлекторном уровне. Существуют сложные задачи, решение которых требует многократного (циклического) обращения к памяти через поле внимания, что соответствует сознательному уровню организации поведения. Это происходит, например, в случаях, когда задачу приходится разбивать на подзадачи. Пример — сложение чисел, когда каждый следующий разряд суммы последовательно проходит через сознание, требуя для своего вычисления новой ассоциации с таблицей сложения. На сознательном уровне решаются задачи, для которых в памяти не хранятся готовые решения, и их приходится искать. Узловые моменты этого поиска, проходя через сознание и вызывая дальнейшие ассоциации, запоминаются, во-первых, для того, чтобы не приходилось повторять уже сделанные ходы, а во-вторых, чтобы найденным решением можно было бы потом воспользоваться на условно-рефлекторном уровне.

В основе сознательного поведения лежит процесс возникновения желаний.

Желание — это осознанное побуждение, т. е. фиксируемое в памяти так, что его можно потом воспринимать, стремление сохранить или избавиться, достичь или избежать какого-либо свойства текущей ситуации. Оно определяет некоторую долговременную (конечную или промежуточную) цель поведения. В отличие от побуждений, величина которых постоянно зависит от текущей ситуации и с ее изменением может резко изменяться, величина желания, будучи занесена в память в виде данных основного типа, служит постоянным стимулом в поведении, т. е. влияет на процесс формирования побуждений до тех пор, пока в результате осознания некоторого побуждения, например, не появится новое желание, противоречащее первому.

Эта общая модель человеческого поведения, на основе которой объясняется феномен мышления и выделяются основные компоненты модели, может служить некоторой основой для рассмотрения того, как модель интеллекта использу-

ет знания и умения, в том числе и лингвистические. Достаточно сказать, что все потоки информации внутри модели между ее составными частями могут быть представлены в символах некоторого языка, не обязательно естественного, но тем не менее образующего знаковую систему, совершенно подобную человеческому языку. Степень сложности передаваемой информации определяет степень сложности используемых языковых средств. Не касаясь здесь теории языкового знака, можно безошибочно утверждать, что символьный язык внутри модели имеет как план выражения, так и план содержания, и что он может изучаться как некоторая самостоятельная сущность.

## 5.3. Представление знаний

Следующим важным этапом рассуждения будет определение того, что именно передается при обмене информацией между компонентами мыслительной структуры. Каковы способы представления «содержания» человеческого мышления, способы представления того, что мы называем «знаниями»?

В общем виде под «знаниями», используемыми в системах моделирования интеллекта, понимаются особым образом организованные данные. Заметим, что здесь возникает некоторый замкнутый круг («знания» = «данные» или «сведения»). Этот круг, однако, не мешает интуитивному представлению о том, что такое знания. Для того, чтобы знания могли использоваться в системе, они должны быть формализованы, для чего применяется специальный математический аппарат. Разные способы представления знаний отличаются друг от друга видом и характером применяемого математического аппарата.

В большинстве случаев в основе представления знаний лежит логическая форма выражения. Элементарной единицей чаще всего выступает суждение или утверждение (Кривоносов, 1996). Знание некоторой системы, определенной предметной области, складывается из совокупности утверждений относительно этой системы или области. Для представления знаний используется аппарат формальной логики, в частности, исчисление предикатов, позволяющий не только фиксировать мно-

жество заданных суждений, но и выводить из них, с помощью специальных правил, некоторые новые суждения.

Обычно выделяют два способа представления знаний: декларативный и процедурный. Декларативное представление знаний есть множество утверждений, в значительной степени не зависимых от того, где их можно использовать. Оно состоит из аксиом, всех выведенных к тому времени теорем (в широком смысле слова) и множества операторов, представленных правилами вывода. Описывая некоторый факт или явление декларативно, мы можем обрисовать его структуру в терминах определенных признаков, указать на связь с другими аспектами или явлениями, назвать связанные с ним прагматические аспекты. Однако в этих знаниях не будет указаний на те реальные действия, которые должны породить эти знания. В этом смысле декларативные описания пассивны. Примером декларативных описаний может служить научная классификация явлений, например систематика растений, основы которой изложены К. Линнеем. Там установлены системы признаков, определяющих различные виды и подвиды растений, указаны родовидовые и классо-элементные отношения, однако какие-либо действия не обозначены. Есть, правда, классификации, в которых факты или явления определены по признаку принадлежности к какой-либо ситуации. В этом случае уже подразумеваются некоторые действия. Д. А. Поспелов отмечает, что в этом случае уже в декларативных описаниях содержится некоторая информация о процедурах (Поспелов, 1983).

При процедурном представлении информации знания даются в виде процедур их преобразования для данной предметной области. Пример процедурного представления знаний — инструкция, которую можно прочитать в кабине междугороднего телефона-автомата. Сведения о признаках сведены тут к минимуму и присутствуют только когда, когда наличие или отсутствие определенных свойств приводит к остановке процедуры (например, признак «юбилейности» или «гнутости» опускаемых в автомат монет). Процедурные описания, подобно декларативным, могут вступать между собой в определенные отношения. Так, сценарии или семантические сети специального типа могут связывать процедуры друг с другом. Свя-

зи эти могут быть иерархическими, ситуативными или другого типа.

Противопоставление декларативного и процедурного способов представления знаний не означает их коренной несовместимости; на самом деле, как видно из сказанного выше, в декларативном представлении есть элемент процедуры, а в процедурном — составляющие декларативного описания, например признаки. Еще большее слияние этих двух способов имеет место в таких представлениях знаний, как, например, семантические сети или фреймы.

### 5.3.1. Семантические сети и фреймы

Семантическая сеть представляет собою граф, в вершинах которого фиксируются наименования данных, а дугам соответствуют отношения, устанавливаемые между ними. Если некоторая предметная область может быть представлена набором двухместных предикатов, группирующихся в блоки, представляющие собой множество утверждений о некотором объекте, то она может быть представлена и в виде семантической сети. Бинарные предикаты допускают представление с помощью графов, где дугам соответствуют предикаты, а вершинам — аргументы. Совокупность таких структур, представленных в виде графа, и образует сеть, называемую семантической. В реальных системах применяются сети различной степени сложности, в некоторых из них дугам могут соответствовать формулы или сложные математические модели.

По мнению некоторых специалистов, семантические сети как способ представления знаний имеют определенные преимущества перед декларативным и процедурным представлениями. Они могут обеспечить достаточно легкое понимание, обновление и усвоение знаний в относительно однородной структуре. Кроме того, они реализуют достаточно простой доступ к знаниям, что непосредственно связано с общей эффективностью работы систем. Упрощается процедура вывода решения. Однако для представления простых логических отношений типа логических связок, кванторов общности и существования семантические сети уступают декларативному пред-

ставлению, а в части изображения динамических и параллельных процессов — процедурному.

Сочетание приемов из декларативного и процедурного представлений с принципами, используемыми в семантических сетях, привело к созданию другого типа семантического представления — фреймов. Их особенностью является то, что здесь используются модульные структуры, представляющие собой некоторые относительно самостоятельные блоки или единицы.

Фрейм можно рассматривать как один из видов воспринимаемого объекта, вид в перспективе, который (вид) может быть представлен как некоторый граф (Шемакин, 1985). Вершина такого графа соответствует наименованию объекта, а подчиненные вершины — элементам этого объекта, которые наблюдатель видит с определенной точки зрения. Изменение положения объекта относительно наблюдателя приводит к формированию других фреймов, поскольку видимыми становятся уже другие элементы объекта. Невидимые элементы не исчезают из памяти, а запоминаются, что находит выражение и в формальной записи новых фреймов. Группа связанных между собою фреймов образует систему.

Таким образом, фрейм — это некоторая структура, содержащая сведения об определенном объекте и выступающая как целостная и относительно автономная единица знания. В долговременной памяти человека хранится большой набор систем фреймов, которые используются, например, при распознавании зрительных образов. Представление о фреймах получило дальнейшее развитие и интерпретацию. Фрейм отождествляется с некоторой стандартной, однотипной ситуацией, включающей в себя конкретное множество отдельных однородных ситуаций. В зависимости от классов ситуаций различают фреймы визуальных образов, фреймы-сценарии, семантические фреймы и т. п. Во фреймах, используемых для систем ИИ, с целью представления знаний, можно выделить несколько уровней вершин или узлов, иерархически связанных между собой. Каждому такому узлу соответствует определенное его описание. Для узлов верхних уровней таким описанием служит наименование элементов

стереотипной ситуации, являющихся общими для некоторого множества конкретных ситуаций. Описание вершин нижнего уровня, называемых терминалами, состоит из названия данного терминала, заполнителя и описания данного заполнителя. Поскольку такие вершины должны соответствовать элементам некоторых конкретных ситуаций, то позиция заполнителей во фрейме обычно остается пустой. Она заполняется при распознавании конкретной ситуации. При выборе конкретного заполнителя учитывается его описание, представляющее собой определенное требование или условие запоминания. В качестве таких условий могут выступать отсылки к другим фреймам — субфреймы.

Таким образом, структура фрейма включает в себя три основных типа данных: понятие (название фрейма), характеристика (название терминала), значение характеристики (заполнители терминала). Можно считать, что во фрейме реализуются некоторые общие принципы, свойственные организации баз данных, где в качестве единиц выделяются объекты, характеристики и их значения, а также семантическим сетям, в которых различаются абстрактный и конкретный уровни. В то же время фрейм имеет и свою специфику.

Каждую конкретную базу данных или семантическую сеть можно рассматривать как некоторую макроединицу, состоящую из микроединиц, в качестве которых выступают наименования данных или вершин сети. Фрейм является самостоятельной единицей, имеющей черты как макроединиц, так и микроединиц. Статус его как макроединицы определяется тем, что он также является сетью. Но эта сеть локализована данными, необходимыми для определенных объектов или ситуаций. Будучи сетью, фрейм включает в себя и микроединицы в виде наименования вершин. Но он сам выступает в роли микроединицы, когда входит в состав некоторой системы или сети фреймов. Следовательно, фреймы представляют собой единицу, промежуточную между микро- и макроединицами. Отличительной его особенностью является также и то, что не все его элементы заданы априорно, как это имеет место, например, в семантических сетях. Он допускает различные варианты заполнения терминалов, что придает этому способу

представления знаний необходимую гибкость и адаптивность.

Рассматривая разные способы представления знаний о мире, нужно отметить, что в целом ни один из них не является сам по себе достаточным, универсальным и позволяющим игнорировать все другие. Выбор того или иного способа зависит от конкретной задачи и предметной области.

## 5.4. Знание как объект моделирования

### 5.4.1. Лингвистический аспект представления знаний

Совершенствование систем общения прежде всего связано с решением вопроса о том, что такое знание, выступающее в них в качестве объекта моделирования. Знание может рассматриваться в психологическом, гиосеологическом, лингвистическом и других аспектах. Рассмотрим лингвистический аспект представления знаний.

А. И. Новиков полагает, что лингвистическое представление знаний лучше всего проявляется при рассмотрении исходного понятия «текст». В тексте находит отражение связь между конкретными элементами знания и конкретными языковыми средствами, используемыми для их выражения. Подход к представлению знаний, базирующийся на анализе текстов, относящихся к определенной предметной области, включает в себя два основных этапа: 1) выделение и эксплицитное представление структуры содержания отдельного текста, 2) объединение отдельных структур содержания конкретных текстов и построение на этой основе общей системы знания соответствующей предметной области.

Эта процедура предполагает участие в ней человека, который, опираясь на свои знания в данной предметной области, а также умение анализировать и понимать текст, может осуществлять эксплицитное представление содержания отдельных текстов и их интеграцию в целостную систему знаний. В рамках данного подхода первостепенное значение имеет конструктивное определение содержания текста, его структурных

единиц, их соотношение между собой, а также с единицами поверхностного уровня.

Текст рассматривается не только как готовый продукт, но также и как процесс. Это объясняется тем, что содержание текста формируется в интеллекте человека как результат осмысления языковых единиц, составляющих текст, и понимания текста в целом. Следовательно, оно может быть выявлено при учете основных закономерностей восприятия и понимания текста, осуществляющегося на разных уровнях его организации и проходящего ряд этапов. Для определения структуры содержания необходимо выявить такие единицы мыслительного плана, которыми мышление оперирует на заключительных этапах понимания. Основная сложность заключается здесь в том, что понимание непосредственно не заканчивается на анализе текста. Мыслительное образование, возникшее в результате осмысления языковых средств, может подвергаться дальнейшему осмыслению уже независимо от данного конкретного текста. Поэтому собственно содержанию текста должно соответствовать такое мыслительное образование, которое, с одной стороны, образуется на наиболее глубинном этапе понимания, а с другой — является непосредственным результатом воздействия на интеллект совокупности языковых средств, составляющих данный текст (Котов и др., 1987).

Такое мыслительное образование можно назвать внутренней формой текста, которая является формой существования его содержания и находится в определенном отношении с внешней формой текста.

Под внешней формой понимается совокупность языковых средств, включающая их содержательную сторону и реализующая замысел автора. Это то, что дано для непосредственного восприятия и что должно быть осмыслено и понято. То, что понимается, составляет внутреннюю форму. Это такое мыслительное образование, которое возникает в интеллекте человека и соотносится с внешней формой не поэлементно, а в целом соответствует всей совокупности данных языковых средств.

Переход от внешней формы текста к его внутренней форме и составляет сущность процесса понимания.

### 5.4.2. Понимание текста

Понимание сопровождается интенсивным преобразованием и перестройкой исходной (внешней) формы текста, осуществляющейся в результате мыслительной аналитико-синтетической деятельности и приводящей к свертыванию текста. Результатом преобразования и свертывания текста является совокупность «контекстных объединителей» (смысловые знаки, опорные пункты), каждый из которых замещает в мышлении определенный фрагмент текста. Преобразование исходной формы текста проходит ряд этапов и заключается в переводе его на «язык» внутренней речи, которая представляет собой центральное звено, где совершается качественный скачок от «внешних кодов» языка к «внутренним кодам» интеллекта. Перевод на «язык» внутренней речи не является конечным этапом перехода к содержанию. Этот переход продолжается и на следующих этапах. Содержание текста соответствует денотативному уровню отражения, т. е. знанию о том фрагменте действительности, о котором сообщается в тексте.

Представление о денотативном характере содержания имеет важное методологическое значение, поскольку позволяет достаточно четко разграничить языковые явления, выступающие в качестве средства формирования содержания, и само содержание, имеющее экстралингвистический характер. Это означает, что анализ содержания должен осуществляться на уровне не слов, фраз и подобных единиц, а таких отрезков текста, которые соответствуют целостным единицам денотативного плана.

Содержание текста представляет собой динамическую модель ситуации, заданную всей совокупностью средств, составляющих внешнюю форму текста. Основной единицей содержания является денотат, понимаемый как мыслительное образование, соответствующее свернутым образам моделей предметов или предметных комплексов, о которых сообщается в тексте. Содержание текста включает в себя не только те денотаты, которые имеют в тексте непосредственное выражение, но и подразумеваемые, имплицитные денотаты.

Динамическая модель ситуации, задаваемая текстом, представляет собой систему денотатов, способ организации которой не базируется на грамматических и логико-композиционных закономерностях текста, а подчиняется логике предметных отношений, связывающих денотаты между собой. Между элементами внешней формы и единицами содержания нет однозначного соответствия и непосредственного перехода, поскольку содержание соответствует в целом всей совокупности элементов внешней формы. Между внешней и внутренней формой существует опосредованная связь, устанавливаемая мышлением в процессе восприятия и понимания текста.

Содержание текста не существует вне процесса его восприятия и понимания, поскольку оно каждый раз формируется в интеллекте партнера по коммуникации как результат воздействия на него соответствующих сигналов, в качестве которых выступают языковые средства. Оно включает в себя новую информацию, которая находится в определенном отношении с репродуктивной информацией.

Процесс перехода от языковых единиц, составляющих внешнюю форму текста, к его внутренней форме осуществляется не только на основе так называемого языкового знания, за счет особого устройства языкового знака, а в результате многократного перекодирования языковых выражений, базирующегося на их осмыслении.

Этот процесс характеризуется следующей совокупностью данных. Основная языковая единица — слово — в качестве своей содержательной стороны имеет лексическое значение, которое понимается как не жестко заданная область его предметной отнесенности. Через лексическое значение слово соотносится с понятием, являющимся одной из единиц системы знания, формирующейся в интеллекте. Опосредованная связь языковых единиц с денотативным уровнем текста осуществляется в процессе перестройки их предметной отнесенности на основе речевого смысла. Речевой смысл представляет собой осознание непосредственно не данного отношения между значениями сочетающихся слов, составляющими семантическое пространство текста. Содержание текста, формирующееся в интеллекте в виде совокупности денотатов, представляет со-

бой результат, соответствующий наиболее глубинному уровню понимания, возникающему под воздействием на интеллект соответствующих языковых средств.

### 5.4.3. Денотативный анализ текста

Денотативный анализ текста завершается эксплицитным представлением его содержания в виде графа денотатной структуры, вершинам которого соответствуют имена денотатов, а ребрам — отношения между ними. Основные этапы, через которые проходит процесс построения денотатной структуры по А. И. Новикову, который (процесс) реализован в соответствующих компьютерных программах в виде действующей системы смыслового анализа текста, следующие:

1. определение имен денотатов в тексте;
2. выделение наименований «ключевых» денотатов;
3. установление внутренних связей каждого ключевого денотата с другими денотатами данного текста;
4. определение предметных отношений между денотатами;
5. перестройка структуры связей между денотатами, задаваемых текстом, в соответствии с предметными отношениями этих денотатов;
6. выявление имплицитных денотатов и их предметных отношений;
7. формирование целостной структуры содержания с учетом предметных отношений денотатов и их места в этой структуре в соответствии с тем, в какой роли (подтемы, субподтемы и пр.) они выступают в тексте.

Определение структуры содержания текста является предпосылкой к решению проблемы представления знания, но не решает ее полностью. Объясняется это тем, что такие структуры представляют собой отдельные фрагменты знания, содержащиеся в конкретном тексте, в то время как требуется построение системы знания, относящейся к определенной предметно-тематической области в целом. Это может быть достигнуто за счет интеграции отдельных денотативных структур в некоторую целостную структуру.

Данная концепция моделирования знаний с целью извлечения смысла их текста поучила практическую реализацию и используется для решения конкретных задач содержательного анализа и связных текстов (Новиков, 2002).

## 5.5. Моделирование обучения языку

В настоящее время огромное количество работ посвящено вопросам применения компьютеров для обучения языкам. Современная языковая ситуация, при которой знание иностранного языка все в большей мере становится залогом социального успеха, требует активизации практических и научных разработок в области методики, теории и практики преподавания иностранных языков и активизации применения технических средств в изучении вообще языков, в том числе и родного. Создано много достаточно эффективных компьютерных программ, помогающих на разных стадиях изучения языка освоить те или иные стороны языковой деятельности. Эти программы опираются на современную высокоразвитую компьютерную технологию, в том числе подразумевающую использование высокоэффективных средств мультимедиа — изображения, звука, цвета, графики, видео — и прочей техники. Надо также отметить, что современные учащиеся все в большей степени привыкают к компьютерному способу общения и усвоения информации, что в немалой степени способствует эффективности обучения.

Мы будем в настоящей работе исходить из моделирования не отдельных языковых навыков или умений, а от моделирования некоторых принципиальных способностей к обучению и освоению; главным образом это касается моделирования умственных способностей ребенка осваивать язык по мере развития. С другой стороны, моделирование процесса обучения также помогает раскрыть закономерности научения компьютера анализу и синтезу текста. Эта проблема ставилась уже достаточно давно, однако до сего времени в части лингвистического научения достигнуто не так уж много.

### 5.5.1. Обучение ребенка языку

Как видно из предыдущего, «обучение» мышлению и сознательному поведению основано на использовании памяти и накапливаемого опыта. Представляет интерес моделирование процесса обучения ребенка языку. Такое моделирование, как представляется, может дать ключ к пониманию основных законов пользования языком в их становлении.

Кардинальное исследование проблем моделирования обучения языку представлено в «Энциклопедии по искусственному интеллекту», вышедшей в 1987 году. Рассмотрим кратко основные подходы к решению данного вопроса.

Можно выделить четыре модели обучения языку. Они не только отвечают на вопрос о том, при каких условиях язык может быть описан моделью, но и на вопрос, так ли модель «изучает» язык, как это делает ребенок (Hill, 1987). В результате модели можно классифицировать по тому, как они учитывают психолингвистические данные. Модели описываются в структурах знаний, которые создаются по мере того, как модель учится языку. Выход модели — ответ ребенка на поступающую информацию. Этот ответ может иметь форму как действия, так и быть вербальным.

Первая модель — CHILD Селфриджа. Структуры данных модели представлены на рис. 13 в виде квадратов на блок-схеме. Структуры знаний, которые при этом строятся, исходят из того, что концепты существуют заранее; структуры знаний присоединяют лексику к концептам. Эти концепты могут иметь, например, форму грамматик зависимостей. Рамки структур зависимостей концептов используются для того, чтобы строить словарь слов и значений со специальными слотами (пустыми местами), которые впоследствии заполняются, и позиционной информацией, которая применяется к слотовым фильтрам для правильного нахождения места в предложении по отношению к глаголам. Модель использует механизмы для фокусировки внимания и набор правил усвоения (обучения), в также правил инференции.

В исходном состоянии модель не имеет языка, и она постепенно учится понимать команды и отвечать на них. Пред-

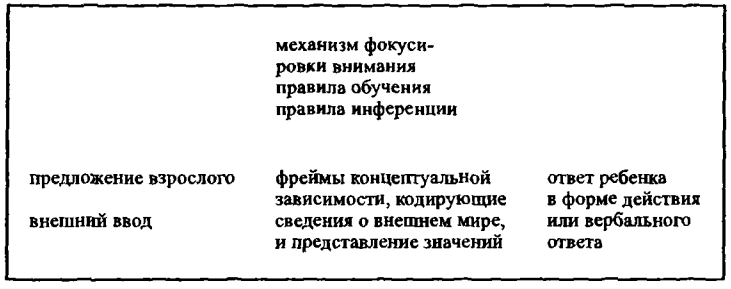


*Рис. 13.* **Компоненты модели Селфриджа CHILD**

ложения взрослого берутся из набора, который получен опытным путем из бесед взрослого с обучаемым ребенком. Понимание выступает в форме некоторого соответствия (mapping) между фреймами концептуальной зависимости и лексической информацией. Модель успешно изучила значительные фрагменты английского и японского языков.

Модель Лэнгли AMBER представляет собой модель, которая обучается языку посредством исправления ошибок. Система Лэнгли показывает процесс постепенного овладения языком во времени и в том порядке, в котором усваиваются морфемы.

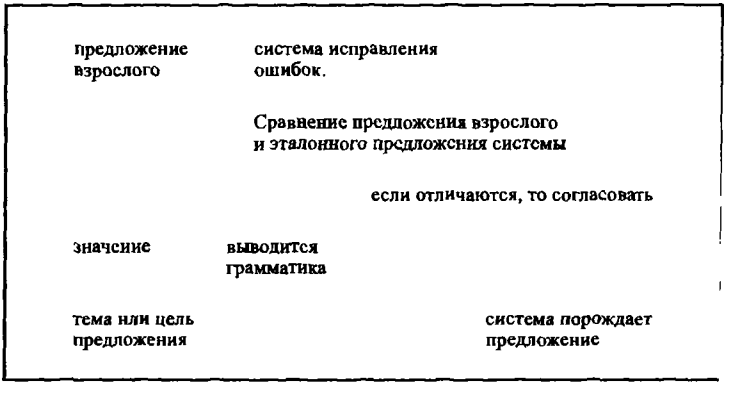


*Рис. 14.* **Компоненты модели Лэнгли AMBER**

Компоненты этой модели показаны на рис. 14. Ввод в модель состоит из порождаемых случайным образом предложений взрослого одновременно с представлением значений и с представлением основной темы предложения. Выход системы — предложение, которое она генерирует.

Система AMBER получает пропозицию, предсказывает предложение и затем сравнивает порожденное ею предложение с поступившим на вход предложением взрослого. Она также устраняет расхождение между этими двумя предложениями, если они имеют место. Цель предложения, порождаемого системой AMBER, — описать основную тему предложения. Модель реализована на языке программирования PRISM, который обеспечивает программную реализацию обучения. Значение представляется древесной структурой, которая использует небольшой набор отношений, таких, как «объект», «агент», «действие», «размер», «цвет». Кроме того, определены такие функции как единственное и множественное число, настоящее и прошедшее время и пр. Через представление значений система понимает и учится «говорить» на естественном языке.

Модель Хилла представлена на рис. 15. Вход в модель — предложения взрослого из специально подготовленного списка. Выход — ответ ребенка в виде предложений, повторяющих или отвечающих на предложения взрослого в соответствии с текущим состоянием грамматики модели. Внутреннее представление модели состоит из динамических структур и физического контекста диалога. Модели сообщается основной лексикон и набор концептов с соответствующим интерфейсом между ними. Не делается никаких предположений относительно конечной формы грамматики взрослого или относительно того, что должно быть построено в модели, но модель постепенно развивается, шаг за шагом, представляя тем самым прогресс ребенка. Процессы, происходящие в модели, обеспечивают слежение за входом и используют специальный набор правил для того, чтобы сосредотачивать внимание на данных, содержащихся в предложениях взрослого с тем, чтобы на их основе формировать классы слов и правила грамматики. Модель реализована на языке программирования LISP и исполь-

зует язык семантических сетей, а также грамматические правила в виде фреймов и лексикон. Модель использует свой языковой опыт для того, чтобы включать слова в классы слов и строить грамматику, которая на первых порах представляет из себя простую фреймовую грамматику, а затем превращается в такую, которая лучше всего может быть описана как набор рекурсивных контекстносвободных правил непосредственно составляющих.

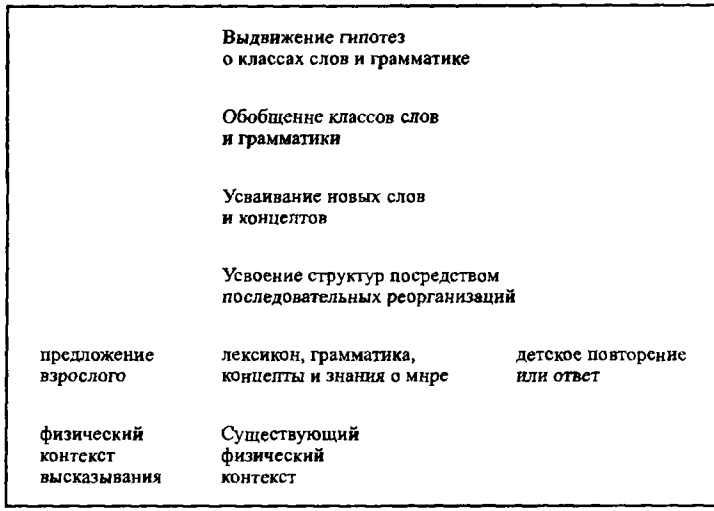


*Рис. 15.* **Компоненты модели Хилла**

В модели Маквинни и Андерсена были учтены данные многих языков. Каждое высказывание взрослого сопровождается пропозиционным представлением цепочки событий, которую оно представляет. Выходом модели является детское предложение. Модель использует парадигму параллельного взаимодействия и соревнования (competition) между структурами данных для того, чтобы породить лексическую функциональную грамматику. Модель реализована на языке программирования Franz LISP. Представление значений в модели принимает форму сложных соответствий между значением и высказыванием. Предполагается, что ребенок пытается учить

слова с помощью тех значений, которые он хотел бы выразить. Поэтому ребенку предоставляется такая форма ментального представления, которая включает пропозиционную структуру, служащую основой для семантической интерпретации. Изучается при этом набор лексических структур, которые реализуют синтаксические правила английского языка.

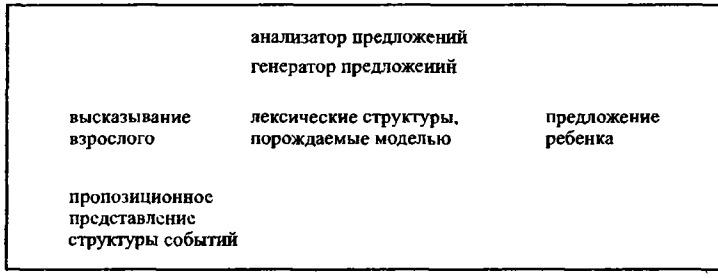


*Рис. 16.* **Компоненты обучающейся модели Маквинни и Андерсона**

### 5.5.2. Сравнение моделей

Оценивая в общем принцип действия и перспективы развития каждой из моделей и вообще моделирования распознавания языка и обучения пользованию языком, можно сказать следующее.

Модель Селфриджа начинает с того места, с которого начинает и ребенок, имея в голове некоторые концепты, но не умея пользоваться языком. Модель начинает действовать как человек, плохо знающий язык, который постепенно накапливает опыт, делает соответствующие ошибки и исправляет их, постепенно улучшая свое понимание. Научившись сначала понимать только слова, модель затем учится понимать синтаксическую информацию и ассоциировать ее со словами. Способность научиться говорить стимулируется способностью научиться понимать, что, благодаря врожденному инстинкту, является стимулирующим фактором. Эта модель может изучать и другие языки, кроме английского, и Селфридж полагает, что его модель может изучить язык до уровня взрослого человека.

Поскольку модель в ее нынешнем виде может воспринимать только простые команды и декларативные предложения, потребуется еще посмотреть, какие могут понадобиться модификации нынешних правил для того, чтобы воспроизвести сложность языка взрослого человека.

Модель Лэнгли начинает с высказываний, состоящих из одного слова, и постепенно учится производить предложения такой сложности, которая характерна для языка взрослого. В отличие от модели Хилла, модель Лэнгли учит суффиксы, префиксы и морфемы. Она не учится пониманию, но получает в готовом виде представления значений для «взрослых» высказываний, которые она в конечном счете научается воспроизводить во всей их сложности. Этой системе не нужны сложные знания о внешнем мире, поскольку она не пытается отвечать на вопросы. Идея грамматики, воплощенной в системе в виде процессов понимания и производства языка, с психолингвистической точки зрения весьма привлекательна.

Модель Хилла начинает с весьма незначительной информации и научается понимать и производить языковые высказывания на уровне двухлетнего ребенка. Эта модель постепенно приобретает хорошие языковые навыки. Она может повторять предложения взрослого в детской манере и отвечать на вопросы. Система устроена так, что подход, с помощью которого изучение языка базируется сначала на чисто психолингвистической основе, а потом переходит к освоению синтаксической системы. Обучающая система становится все более и более лингвистически самостоятельной и не зависящей от своего семантического и психолингвистического происхождения. Именно таким путем ребенок превращается во взрослого. Однако эта модель никогда не переходит уровня двухлетнего ребенка. Эта модель в большей степени, чем другие, предоставляет теоретическое обоснование для лингвистического экспериментирования и обладает большей гибкостью, включая различные параметры и альтернативные модули, которые могут добавляться к модели или включаться в нее с тем, чтобы наблюдать разные результаты.

Модель Маквинни также начинает с высказывания в одно слово и затем приобретает все более сложные навыки владения

языком. Чтобы получить полноценные предложения на выходе, модель должна использовать тщательно подготовленное и расширенное представление значений. Предполагается, что при наличии достаточно полной информации модель сможет решать даже такие сложные проблемы, как анализ конструкций с включением и относительные предложения.

Хотя ни одна из моделей не использует какого-либо специфического набора лингвистических универсалий, в то же время ни одна из них не противопоставлена теориям лингвистических универсалий, и можно предположить, что любая из таких теорий может использовать некоторые врожденные умения и процессы. В вычислительной модели необходимо точно знать, что в ней заложено, и уметь воспроизвести ее конструкцию. По мере того как в будущем будут создаваться более совершенные модели, можно будет точно утверждать, каких именно сведений не хватает для совершенствования моделей.

Из вышеизложенного можно понять общий подход к конструированию таких моделей. Подробности их строения и программной реализации обычно редко публикуются и могут быть понятны только специалистам в данной узкой области. Однако для нашего рассмотрения ясно, что моделирование идет по линии воспроизведения основных различающих функций человеческого владения языком: от общей постановки задачи через словарь и исходный набор грамматических конструкций ко все более самостоятельному овладению основным массивом синтаксических правил построения осмысленных высказываний.

Подробно современные обучающие модели проанализированы в книге Р. К. Потаповой (Потапова, 2002). Она определяет образовательную технологию как способ реализации содержания обучения, предусмотренного учебными программами, представляющий систему целей, форм, методов и средств обучения. В последние годы в системе образования России в соответствии с концепцией информатизации образования усиленно развивались следующие основные направления работы по созданию и внедрению перспективных информационных образовательных технологий:

1) применение универсальных информационных технологий в учебном процессе;

2) применение компьютерных средств телекоммуникаций;
3) разработка и использование в учебном процессе компьютерных обучающих и контролирующих программ, компьютерных учебников и т. п.;
4) разработка и использование мультимедийных программных продуктов, содержащих значительный объем информации, например «электронных энциклопедий», сопровождаемых методическим материалом (набором учебных заданий) для их использования в целях обучения.

## 5.6. Моделирование на уровне статистики. Квантитативная лингвистика и связанные с ней дисциплины. Теория и результаты

Статистика является мощным инструментом для выявления сущности лингвистических явлений. Применение статистики в языкознании имеет свою историю, несмотря на то, что многие филологи и лингвисты до сих пор считают, что язык противопоставлен мере и числу. Прикладная лингвистика и компьютерная лингвистика, предметом которых являются практические приложения, базируются не только на теоретических постулатах, но и на изучении массивов текстов, и в таком изучении не представляется возможным обойтись без меры и числа.

### 5.6.1. Комбинаторная и квантитативная лингвистика

Выбор математического аппарата в лингвистических исследованиях зависит от того, как определяются основные задачи языкознания. Если считать, что основной задачей должно быть изучение грамматики, порождающей текст, причем грамматика понимается как конечное множество детерминированных правил, а язык рассматривается как бесконечное множество регулярных цепочек слов, порождаемых такой грамматикой, то тогда экспликация лингвистических явлений

должна опираться на такие разделы неколичественной математики, как теория множеств, математическая логика, теория алгоритмов и т. д. На основе применения этого аппарата сложилось направление комбинаторной лингвистики. Однако только этим трудно объяснить все многообразие лингвистических явлений и языка. Многие выдающиеся ученые неоднократно обращали внимание на полезность и важность квантитативного подхода к изучению лингвистических объектов. «Нужно чаще применять в языкознании количественное, математическое мышление», — указывал Бодуэн де Куртенэ. Академик В. В. Виноградов, говоря о частотности разных типов слов в разных стилях книжной и разговорной речи, отмечал, что «точные изыскания в этой области помогли бы установить структурно-грамматические, а отчасти и семантические различия между стилями». «Частотность принадлежит функциональной стороне языковой системы ... учет частотности любого языкового явления — полезный прием при анализе» (В. Н Ярцева) (Цит. по Тулдава, 1987). Полезный и интересный материал о статистическом моделировании в языкознании содержится также в разделе «Методы статистического моделирования в языкознании» учебника «Прикладное языкознание» (Мартыненко, 1996).

Квантитативные методы не в состоянии решать любые проблемы анализа языковых явлений. Квантитативный подход способен охватить лишь определенный аспект языка и речи. Но это — существенный аспект, отражающий ряд важных сторон речевой деятельности, которые невозможно обнаружить чисто качественным анализом.

Всякая количественная характеристика лингвистических явлений предполагает их качественную характеристику. В то же время качественная характеристика лингвистического объекта существенным образом зависит от количества элементов, его образующих, от частоты их употребления или от силы взаимодействия (корреляции) этих элементов. Можно констатировать наличие тесной взаимосвязи качественных и количественных характеристик языка: совместное их рассмотрение открывает широкие лингвистические возможности исследования языковых процессов и явлений.

### 5.6.2. Язык и речь

Для квантитативно-системного изучения языковых явлений важно разграничение языка и речи. Можно полагать, что есть два взаимосвязанных, но отделимых друг от друга компонента: средство (орудие) и его применение. Комплекс «язык-речь» можно представить в виде перекрещивания двух главных осей: оси с противопоставлением потенции и реализации и оси с противопоставлением динамики и статики языка (по Тулдаве, 1987). В этой системе фигурируют язык и речь.

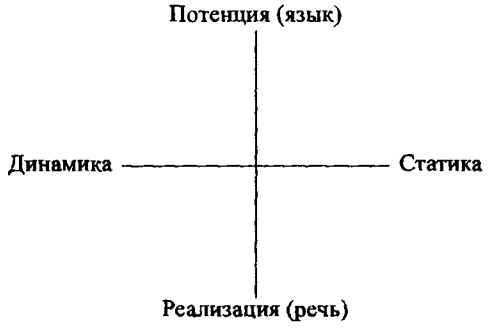


*Рис. 17.* **Язык — речь, потенция — реализация**

В соотношении потенции-реализации содержится элемент уровневости: потенция — высший, а реализация — низший уровень. Однако в онтологическом плане наоборот: речь предшествует языку, речь как конкретная реализация языка является единственным непосредственно наблюдаемым объектом лингвистики.

В квантитативной лингвистике противопоставление языка и речи имеет прямой практический смысл. С потенцией и реализацией связана идея о «полной системе», т. е. о полной группе событий, которые теоретически могут произойти в данных условиях, в противопоставлении к ограниченному набору действительно реализуемых событий. Можно связывать потенцию и реализацию с соотношением между генеральной совокупностью и выборкой из этой генеральной совокупности: первая относится к языку, вторая — к речи. С потенцией и реализаци-

ей можно связывать понятия вероятности и частотности: язык вероятностен, речь — частотна (по Б. Н. Головину).

Характеристики динамики и статики используются часто при разграничении речи как процесса и речи как результата этого процесса. Однако уровень языка рассматривается в чисто статическом плане как «инвентарь языковых средств и набор правил». Разницу между сферами динамики и статики видят еще и в том, что динамика (механизм и процесс) связывается с деятельностью мозга, а статика (язык как предмет) находится вне человека.

### 5.6.3. Аспекты речевой деятельности

В итоге в системе речевой деятельности можно различать четыре основных аспекта (подсистемы): на языковом уровне — языковую компетенцию и языковую схему, на речевом уровне — речевой процесс (акт) и речевой продукт (текст).

| *Аспекты речевой деятельности* | | |
|---|---|---|
| Признаки | Динамика | Статика |
| Потенция (язык) | Языковая компетенция | Языковая схема |
| Реализация (речь) | Речевой процесс (акт) | Речевой продукт (текст) |

*Рис. 18.* **Речевая деятельность и ее аспекты**

*Языковая схема.* Под ней понимается система языковых элементов и отношений между ними. Эти отношения могут относиться как к синтагматике, так и к парадигматике. Языковую схему можно рассматривать как статическую систему данного языка в целом, т. е. как суммарную совокупность лингвистических элементов и отношений между ними. Языковая схема отличается устойчивостью в рамках данного языка, в ней выделяется, например общая лексика, присущая всем подъязыкам, и она характеризуется стабильными связями и закономерностями на общеязыковом уровне. Выработавшееся в общественной практике типичное и общепринятое языковое употребление, регулярно повторяющееся в определенной сфе-

ре коммуникации и отражающееся в языковой схеме в виде устойчивого ядра подсистемы, называется *нормой* для данного языка. В квантитативной лингвистике норму можно определить как наиболее вероятный состав элементов и наиболее вероятные отношения между элементами.

*Языковая компетенция.* Языковая компетенция является следствием отражения сознанием людей языковой схемы, т. е. это набор элементов и системных отношений между ними, сходных в плане отражения с элементами и отношениями языковой схемы, плюс особый динамический компонент, необходимый для приведения языка в действие. Без обращения к соседним наукам лингвистика не в состоянии объяснить действительную природу и внутренние закономерности речевой деятельности в целом. Точно так же квантитативные закономерности речи (универсальные особенности статистической структуры текста, наличие устойчивых распределений и регулярности в корреляции с внешней средой и пр.) могут быть объяснены только при условии привлечения к анализу экстралингвистических представлений и критериев.

*Речевой процесс.* Речевой процесс является реализацией потенции языка, т. е. самим действием механизма порождения речи. Отношение речевого процесса к речевой компетенции можно рассматривать также как отношение управляемой системы к управляющей. Речевой процесс имеет своим непосредственным результатом линейную последовательность речевых единиц. Сам процесс порождения высказываний представляет собой сложный многоступенчатый речемыслительный акт, состоящий из этапов превращения исходного замысла через внутреннюю речь в схему речевого высказывания и включения его в фонематические, лексико-семантические и логико-грамматические коды языка (по А. А. Леонтьеву).

*Речевой продукт.* Это результат процесса порождения речи, или текст, который в самом общем виде понимается как определенным образом (обычно письменно) зафиксированный отрезок речевого континуума. По отношению к речевому процессу текст характеризуется тем, что в нем снята процессуаль-

ная форма речемыслительной деятельности, хотя он и сохраняет связь с этой деятельностью.

Будучи единственным прямо наблюдаемым объектом лингвистического исследования, текст при соответствующем подходе может служить моделью изучения также и динамических сторон речи.

Учитывая, что порождение речи рассматривается как сложный вероятностный процесс, формирующийся в результате взаимодействия детерминированных и случайных факторов, можно сделать вывод, что и результат такого процесса неизбежно должен характеризоваться вероятностными свойствами. В целом текст и соответствующий словарь могут рассматриваться как вероятностные системы. Эмпирически это проявляется, с одной стороны, в наличии стабильных распределений единиц, в устойчивых корреляциях внутрисистемных и межсистемных образований, в образовании «ядра» в замкнутых группах, а с другой стороны, — в явлениях периферии, размытости границ и различного рода флуктуациях.

Таким образом, мы видим, что квантитативная лингвистика и приемы математической статистики позволяют нам получить объективные данные лингвистического характера для решения прикладных лингвистических задач — построения словарей, алгоритмов автоматического анализа и синтеза, систем смыслового и содержательного анализа текстов на естественном языке.

Основными понятиями квантитативной лингвистики являются понятия случайной выборки и совокупности, а также частоты встречаемости и распределения частот встречаемости единиц разного уровня в текстах. На основе законов Ципфа и Ципфа-Мандельброта построена теория ранговых распределений, которая широко применяется для анализа текстов. В фонетике методы квантитативной лингвистики нашли свое применение для построения системы фонем на основе критерия частоты встречаемости, для разработки квантитативных моделей слогоделения. Продуктивен также метод дистрибутивно-статистического анализа, с помощью которого можно получить весьма точные и полезные данные о лексическом составе тек-

стов. Построение частотных словарей, конкордансов, создание машинных фондов лексики невозможно без применения методов квантитативной лингвистики. Грамматический и семантический аспекты квантитативного исследования лексики привели к построению математической теории словообразования, определению лексико-семантических групп. Такие современные методы анализа текстов, как кластерный анализ, контент-анализ, лексико-стилистический анализ также используют достижения квантитативной лингвистики (Максименко, 2002).

# *Глава 6*
# Экспертные системы

Экспертные системы в настоящее время занимают довольно много места в общей проблематике искусственного интеллекта. Под экспертными системами понимаются автоматизированные системы информации, выполняющие роль консультантов (Шемакин, 1985). Каждая экспертная система ориентирована на определенную предметную область, например политику, медицинскую диагностику, проектирование технических систем и пр. Экспертная система хранит знания многих специалистов. При разработке в систему закладываются начальные знания, при обучении они обобщаются. Система, получив на входе задачу, анализирует ее и строит план решения.

Любая экспертная система включает базу данных, состоящую из правил, которые при функционировании системы используются в традиционных интеллектуальных операциях поиска по образцу и синтаксического отождествления.

## 6.1. Способы организации знаний в машине

Традиционные способы работы со знаниями, хранящимися в экспертных системах, основанные на их формальном представлении, мало пригодны из-за сложности описания самих предметных областей. Поэтому поиск удовлетворительного решения идет в направлении гуманизации машинных знаний, суть которой состоит в отыскании такого изображения для машинного представления знаний, которое выражало бы их в понятных и привычных для человека терминах. С этой точки зрения актуальной задачей проектирования современных автоматизированных систем обработки информации является задача научить систему понимать тексты на естественном языке.

Часто лингвистическую, языковую часть экспертных систем называют лингвистическим процессором. Ядром лингвистического процессора является семантический анализ предметных проблем, т. е. анализ системы предметных объектов конкретной области и связей между ними, образующих тематическую область. По мнению некоторых специалистов (см., например, Козловский, 1986), наиболее эффективным семантическим анализом в экспертных системах на сегодня является такой, в котором реализуются причинно-следственные связи между двумя явлениями, хотя в общем случае экспертная система может реализовать любые виды связей — ассоциативные, по сходству и подобию и пр., необязательно описываемые в логической форме. Этим, собственно, экспертная система отличается от информационно-логической системы, где новые знания выводятся из уже имеющихся в машине посредством логических операций.

Для успешного применения лингвистического процессора совместно с экспертной системой эта система должна иметь определенные свойства. База знаний такой системы должна описывать: 1) свойства предметной среды, например заболевания или технические устройства — предметную область; 2) грамматику языка, который используется для коммуникации; 3) коммуникативную систему, которая включает пользователей, экспертную систему и, возможно, другие системы.

## 6.2. Современные экспертные системы

Знания в экспертной системе обычно представляются в следующих формах:

- декларативная форма, содержащая представления в виде структур объектов, в частности, описывающая причинно-следственные отношения явлений;
- процедуральная (процедурная) форма, включающая правила организации, отслеживания, удаления отношений и структур;
- инструментальная форма знаний, включающая средства построения и преобразования как структурной, декларативной, так и процедуральной формы.

Инструментальная часть играет роль ассемблера и базового ядра, с которого начинается развитие системы. В него может входить один или несколько развитых языков программирования и транслятор или интерпретатор для построения и исполнения декларативной и процедуральной форм представления знаний.

Реализация экспертных систем показывает, что процедурная часть занимает сравнительно небольшой объем и чаще всего представляет собой систему продукций (о продукциях см.: Шемакин, 1985). Гораздо больше проблем возникает в связи с декларативной частью базы знаний, которая, как правило, огромна. Вследствие этого имеют место частые обращения к запоминающим устройствам и возникают сложные алгоритмы переработки этой информации. Можно предполагать, что широкое применение систем такого класса наступит с распространением персональных ЭВМ достаточно большой памяти и быстродействия.

Экспертная система представляет собой современное хранилище знаний. Это хранилище идет на смену традиционным формам представления и хранения информации, знаний и сведений. Одной из таких форм является хорошо известная нам книга. Экспертную систему можно рассматривать как некоторую электронную книгу.

Сравнивая обычную книгу и современную электронную книгу, В. Н. Агеев отмечает, что, рассматривая эволюцию средств общения человека, можно увидеть много общего между первым (довербальным) и современным (текущее столетие) периодами: современные технические средства имеют ярко выраженную образную направленность (звуки, музыка, свет, движение, зрительные образы); т. е. в процессах общения стали преобладать те знаковые средства, которые появились еще на заре развития человечества и которые с появлением письменности и развитием книгопечатания были заметно потеснены абстрактно-логическими типами знаков.

Существует несколько основных типов знаковых систем, порождающих все многообразие используемых в настоящее время искусственных, естественных и эволюционно-развивающихся языков. Это: 1) натуральные или естественно-знаковые

системы; 2) иконические знаковые системы; 3) системы художественных образов; 4) системы речевых знаков; 5) системы письменных знаков; 6) формализованные или кодовые знаковые системы. В этом ряду каждый последующий тип отличается от предыдущего более высоким уровнем абстракции используемых знаков, а также все более возрастающей ролью соответствующего им метаязыка. С точки зрения науки эргосемиотики, занимающейся вопросами взаимодействия человека с компьютером, электронная книга предстает перед нами как обобщенная абстрактная система физических и знаковых средств и как конкретная совокупность этих средств, позволяющих пользователю динамично и творчески взаимодействовать с информационным массивом. Это многоуровневая информационная система, создаваемая на основе разнообразных естественных и искусственных языков, а также языков, имманентно присущих человеку (Агеев, 1996). В качестве примера таких электронных комплексных книг можно назвать компьютерные энциклопедии «Кирилл и Мефодий» российского производства или The British Multimedia Encyclopedia, Encarta Encyclopedia и др.

Чрезвычайно интересной является постановка вопроса о связи живой природы с текстами. А. Е. Седов, рассматривая понятие текста как «последовательности знаков, построенной по правилам системы того или иного языка», отмечает, что тексты есть как в феноменах культуры (в различных видах письменности), так и во всех живых организмах (в генетических программах). Именно благодаря тесным связям между понятиями «жизнь» и «текст» еще задолго до открытия биогенных текстов — последовательностей нуклеотидов в ДНК и РНК и аминокислот в белках — в разных формах запечатления знаний формировались познавательные модели типа «жизнь как текст». Именно такие модели могли способствовать разным открытиям. Ж. Лежен высказал предположение, что даже абстрактные математические объекты (множества, графы и пр.) — это изоморфные отображения реальных субстратных структур, формирующихся в мозгу. Сейчас эти идеи подтверждают нейробиологи и когнитологи и воплощают создатели виртуальной реальности. Рассматривая далее роль текстов в

биологии, А. Е. Седов высказывает интересные мысли о роли текста в систематизации и хранении знаний не только с точки зрения чисто знаковой, но и с точки зрения самого физико-химического существа процесса мышления. Надо полагать, что такая точка зрения может оказать влияние на моделирование процессов хранения и выдачи информации в виде экспертных систем (Седов, 1998).

Большое значение приобретают в современном информационном мире экспертные обучающие системы. Анализируя современный инвентарь экспертных систем, Р. К. Потапова отмечает, что важнейшая роль в экспертных обучающих системах отводится базам знаний. База знаний должна реагировать на действия обучающегося. Она сама должна «решать задачи» по мере надобности в той предметной области, которой программа обучает. Процесс решения задачи может быть показан или не показан обучающемуся. Разработаны два типа баз знаний: эксперты «черный ящик» и «прозрачный ящик». Эксперт «черный ящик» осуществляет алгоритмический подсчет для принятия оптимального решения в данной предметной области. Знания и процесс принятия решения представлены для обучающегося имплицитно. Эксперт «прозрачный ящик» не только решает задачу, но и показывает обучающемуся, каким образом был достигнут правильный результат.

В экспертных системах функционируют две главные формы представления знаний: факты и правила. Факты фиксируют количественные и качественные показания объектов и процессов (действий). Правила описывают отношения между фактами в виде логических условий, связывающих посылки и следствия (Потапова, 2002, с. 36).

# *Глава 7*
# Машинный перевод как центральная проблема искусственного интеллекта

## 7.1. Значение идеи машинного перевода

### 7.1.1. Машинный перевод и теория языка

Идея машинного перевода, т. е. мысль поручить машине работу по переводу с одного естественного языка на другой, насчитывает к настоящему времени уже около пятидесяти лет существования. Примерно столько же лет ведутся научно-исследовательские и опытно-конструкторские работы по машинному (автоматическому) переводу во многих странах мира. Параллельно с созданием систем машинного перевода (МП) разрабатывались и разрабатываются теории, которые можно было бы положить в основу создания таких систем. Интересно отметить, что такие теории, вследствие сложности задачи МП, о чем мы будем говорить далее, как правило, охватывают не только в собственном смысле перевод с одного естественного языка на другой, но и многие смежные актуальные и часто глобальные теоретические вопросы человеческого владения и пользования языком.

Интенсивность работ по МП в разные периоды времени различная, Нынешний период характеризуется некоторым спадом интереса к проблеме, по сравнению, например, с шестидесятыми годами прошлого века. Мода на машинный перевод прошла, многие его энтузиасты давно занялись другими делами. Современное состояние машинного перевода, если его характеризовать несколько критически и в сопоставлении с тем, что имело место ранее, характеризуется некоторыми вялотекущими научными исследованиями, существованием определенного количества практически работающих и про-

дающихся на рынке программного обеспечения систем МП в разных странах и появляющимися время от времени сообщениями о разработке новых систем.

Проблема автоматизации перевода не теряет своей актуальности вследствие того, что: а) перевод с одного языка на другой — единственный эффективный способ преодоления языковых барьеров, поскольку другие (внедрение и использование универсального языка типа эсперанто, изучение иностранных языков основным составом коммуникантов и пр.) неэффективны и не могут заменить перевода; б) растут и расширяются возможности современной компьютерной информационной технологии, поэтому всегда появляется соблазн поручить машине какую-нибудь интересную интеллектуальную задачу; в) спрос на переводы в мире увеличивается в абсолютных и относительных пропорциях соответственно тому, как все больше естественных языков приобщается к мировой цивилизации и вступает в коммуникационную информационную сферу. Высока также и научная привлекательность проблемы машинного перевода.

Чем интересен машинный перевод для теории языка?

Если вспомнить появление самой идеи МП и связанных с этим сенсаций, когда компьютеры, только что появившиеся в научной практике, были окутаны ореолом какой-то тайны и с ними общались только посвященные, то приходится признать, что эта идея произвела огромное влияние на лингвистику. Появилось множество всяких новых лингвистик: математическая, статистическая, алгоритмическая, вычислительная, инженерная и другие. Все они отражали новое направление в языке, новые взгляды на объект, предмет, методы и результаты теоретических изысканий в области естественного языка. Возникли предположения о возможности математического моделирования языка, стал оспариваться тезис о том, что язык внеположен мере и числу, ученые вернулись к тем формальным поискам, точнее, поискам формализмов в языке, которые были известны из трудов выдающихся лингвистов, таких, как Л. В. Щерба, который своей «глокой куздрой» обратил внимание на значение формы в языке, до того находившейся в некотором загоне из-за преобладания семантических взглядов.

Основной смысл новых лингвистик заключался в попытках формализовать разные языковые уровни. Привлекательность машинного перевода для теории именно и заключалась в том, что все эти уровни были разные: задача машинного перевода требует распознавания графических образов (на этапе ввода в компьютер), морфологического анализа, анализа и перевода лексики, слов и словосочетаний, синтаксического анализа и синтеза и, наконец, семантических преобразований, долженствующих обеспечить смысловое равенство введенного и выведенного предложения и/или текста в целом.

Полученная в результате анализа информация должна быть необходимой и достаточной для синтеза переводящих единиц на каждом из уровней.

В этом, т. е. во всеохватываемости, включенности всех уровней языка в задачу и заключается, по-видимому, первая теоретически привлекательная особенность машинного перевода.

Вторая особенность в том, что перевод — единая задача, пронизывающая все уровни языковой структуры. Анализ и синтез единиц всех уровней служат одной конкретной цели — передаче смысла, содержания высказывания (в широком смысле) с одного естественного языка на другой. Разные языковые уровни можно изучать с разными целями. При этом предмет исследования часто становится неопределенным. Теоретикам языкознания и преподавателям лингвистики хорошо известно, что единого курса языкознания нет и быть не может. Каждый профессор читает свой собственный курс языкознания, исходя из собственного понимания основополагающих сущностей этого предмета. Если математика, химия, физика, геометрия и другие точные науки исходят из нескольких фундаментальных теорем или посылок, на основе которых возникают все дальнейшие построения, то в языкознании все зависит от научной школы, в рамках которой определяются исходные понятия: что принимается за фонему, морфему, как определяется слово или словосочетание, исходит ли автор из концепции частей речи или дистрибутивных классов слов, является ли объектом лингвистики (по мнению данного ученого) только предложение и фраза, или же связный текст целиком и т. п. и т. д. «Внеположенность» языка мере и числу затрудня-

ет квалификацию объектов языкознания в терминах точных наук, что и приводит к множественности теорий языка. Задача же формализации и алгоритмизации для перевода заставляет выбирать некоторые сквозные единицы описания уровней языка, такие, на которых можно было бы строить алгоритмы анализа и синтеза не только отдельных уровней, но и всей системы, с расчетом на соответствующий синтез выходного предложения и текста в целом. Определяющим становится некоторый функциональный подход, налагающий ограничения на безграничные фантазии и «полет творческой мысли», не подкрепленный обратной связью инженерных проверок.

Для добросовестного исследователя в машинном переводе всегда открыта возможность проверить правильность теорий и концепций посредством практики. Сделал алгоритм синтаксического анализа — проверь на материале новых предложений, не учтенных при его составлении, как данный алгоритм работает, какие результаты выдает. Сама задача МП, а именно получение выходного предложения, соответствующего входному, есть уже реализация принципа обратной связи. Невозможность проверить правильность заложенной концепции можно, видимо, считать признаком теоретического или практического дефекта системы. Если в начале работ по машинному переводу многие чисто теоретические концепции нельзя было проверить, по мнению их авторов, вследствие недостаточной памяти или быстродействия тогдашних ЭВМ, то теперь такой аргумент отпадает вследствие того, что современные компьютеры имеют значительные объемы памяти и быстродействия, практически достаточные для любой объемной лингвистической задачи.

Третья особенность машинного перевода — это связь его с лексическим уровнем языка. Примитивный взгляд на перевод — а ведь именно он лежал в основе самой идеи машинного перевода, выдвинутой, как известно, математиками и инженерами-программистами, — состоит в том, что нужно переводить слова и словосочетания. Такой взгляд является следствием представления о языке как о некотором коде в прямом, криптографическом смысле слова. Довольно скоро представление о языке как коде было отвергнуто как недостаточное для

описания природы языка, но интерес к лексическому уровню, что весьма естественно, остался. Все-таки значительная часть содержания высказывания заключается именно в словах, не только и не столько в их порядке и согласовании (Комлев, 1992). Без перевода слова и его окружения не может быть проведен анализ предложения, без исследования поведения слова в тексте не может быть вообще никакого перевода, не только машинного. Лексический уровень имеет особое отношение к семантике, а передачу смысла мы на интуитивном уровне считаем главной задачей перевода. Как бы то ни было, лексика для МП играет большую роль, и именно она объясняет разного рода ответвления от МП в виде автоматических словарей в помощь переводчику, терминологических банков данных и пр.

Четвертая особенность МП — на другом конце лингвистического спектра. Операция перевода с одного языка на другой является сложнейшей задачей интеллекта, поскольку требует не только замены слов одного языка словами другого, но и передачи мысли в полном ее объеме, со всеми ее оттенками и коннотациями, будь то политический, научный или технический текст. О машинном переводе художественного текста во времена начальной эйфории по поводу МП говорили только крайние энтузиасты. Тем не менее и в более точных типах текста и стилях, вроде научного и технического, немалое количество так называемых переводческих трудностей, связанных с формулировкой мысли в двуязычной ситуации. Как ни определять мысль, мышление, понимание, восприятие смысла и содержания, ясно, что машинный перевод, поскольку он «перевод», должен передать это содержание на уровне текстовых единиц. Отсюда следует, что задача машинного перевода есть задача искусственного интеллекта, какое бы содержание ни вкладывалось в понятие «искусственный интеллект».

Эти четыре особенности задачи машинного перевода, как моделирования процесса перевода для компьютера, представляли ранее и представляют сейчас фундамент, на котором базируется интерес к машинному переводу, и этот интерес не является преходящим. См. также многие современные работы на эту тему, например (Miram, 1998; Tsujii, 1997) и др.

Почему эту проблему можно считать центральной в искусственном интеллекте?

По нашему мнению, центральным в искусственном интеллекте является вопрос моделирования деятельности мозга как орудия мышления. Мышление неразрывно связано с языком. Связь языка с мышлением никто никогда не отрицал, независимо от того, какие роли отводились каждому компоненту такой связи. Язык — непосредственная действительность мысли. Моделирование языковой деятельности на уровне перевода затрагивает, как показано выше, все уровни языковой структуры, от морфологии до семантики. Поэтому задача машинного перевода как задача моделирования сложнейшего комплексного вида языковой деятельности, есть первостепенная задача искусственного интеллекта. Добавим сюда также и то обстоятельство, что машинный перевод как воспроизводящая инженерно-лингвистическая модель, обладает таким компонентом, как обратная связь, что позволяет оперативно корректировать исходную систему гипотез (Пиотровский, 1979, 1975 и др.)

### 7.1.2. К истории машинного перевода

История машинного перевода чрезвычайно интересна и поучительна. Машинный перевод (МП) начался с появления компьютеров и с попыток применения их к решению интеллектуальных задач. Первыми специалистами по МП были ученые разных отраслей знания, среди них главным образом математики, программисты, теоретики той области знаний, которая впоследствии стала называться искусственным интеллектом. Первоначальной точкой зрения на МП была концепция языка как некоторого кода, и проблема перевода приравнивалась к проблеме перекодирования из одного языка (одного кода) в другой. Однако достаточно скоро стала ясна однобокость такой точки зрения. После этого начались углубленные изыскания в теорию языка, в попытках главным образом найти способы передачи смысла.

История ранних исследований в области МП хорошо отражена в аналитическом труде и сборнике воспоминаний ветеранов машинного перевода, выпущенном Дж. Хатчинсом в

2000 году (Early Years In Machine Translation, 2000). Он включил в сборник всю историю МП, начиная с изобретения русского ученого Смирнова-Троянского, который предложил механического переводчика, автоматически подбирающего словесные эквиваленты для единиц входного языка, до широкомасштабных исследований по МП, которые во второй половине XX века развернулись практически во всех странах и привели к возникновению компьютерной лингвистики, а также дали значительный стимул к развитию общей лингвистической теории, в частности и особенно тех направлений языковедческой науки, в которых главное внимание уделялось формальным и количественным характеристикам языка и речи.

В советской лингвистической традиции исследования по машинному переводу периода 1949–63 годов отражены в библиографии (Мельчук и др., 1967). Современный исследователь найдет в истории машинного перевода много интересных идей, некоторые из которых просто не могли быть осуществлены в то время из-за слабых возможностей компьютеров, а другие показали свою практическую неэффективность в том, что касается собственно МП, но могут представлять интерес с других точек зрения и в свете новых возможностей информационных технологий. Как и всякая история, без изучения того, что было, трудно ожидать успеха в новых разработках.

Периодизацию истории машинного перевода можно строить по разным основаниям. В наибольшей степени отражающей реальность можно считать концепцию следующих периодов истории МП:

*1) 1945–1956 годы*

Первое десятилетие в истории МП знаменуется энтузиазмом по поводу появления новой интересной идеи. Эта идея поддерживается ростом компьютерных технологий и их возможностей в решении интеллектуальных и математических задач, ранее доступных только человеку. В части собственно МП этот период характеризуется взглядом на язык как на код и особым вниманием к составлению машинных словарей, поскольку предполагалось, что перевод слов уже даст возможность понять текст. К этому периоду относится создание первых систем машинного перевода за рубе-

жом и в СССР — Джорджтаунская система в США, ее многочисленные варианты, система МП И. К. Бельской в Институте математики Академии наук СССР, система АМПАР в Министерстве электропромышленности СССР, семейство практических систем машинного перевода в Ленинграде под руководством профессора Р. Г. Пиотровского и ряд других систем. В это же время появляются новые отрасли лингвистической науки и открываются отделения машинного перевода в целом ряде вузов нашей страны, а также в зарубежных университетах.

2) *1956–1966 годы*

Второе десятилетие характеризуется некоторым разочарованием в реальном качестве машинного перевода. Первые системы дают не очень качественный перевод, требующий большого объема редактирования. Затраты на разработку систем практического направления весьма велики, а результаты разочаровывают. Этот период характеризуется ростом внимания к теоретическим проблемам МП — разрабатываются алгоритмы анализа и синтеза разных уровней языковой системы для разных языков. Продолжается и совершенствование действующих систем, причем на первый план выходит моделирование перевода на уровне переводных соответствий данной языковой пары.

3) *1967–1975 годы*

Третье десятилетие характеризуется всесторонним развитием теории машинного перевода, к проблематике которого подключилось за это время предметное поле автоматизированного информационного поиска, что потребовало разработки методов содержательного анализа текстов и построения естественно-языковых интерфейсов. Значительный прогресс в компьютерных технологиях и потребности в автоматизации перевода вновь заставили практиков и теоретиков вернуться к разработке практически пригодных систем МП.

Начиная с этого периода, исследования и разработки в области машинного перевода вступили в фазу постепенного развития, в которой нет уже элемента сенсационности и погони за выдающимися результатами, а есть систематическое иссле-

дование принципиальных положений теории и одновременно стремление решить ряд практических задач информационного обслуживания научно-техническими переводами. Наряду с теоретическими исследованиями и новыми концепциями моделирования перевода появляются действующие системы, дающие практически пригодные результаты для ограниченных областей науки, техники и других подъязыков естественных языков мирового сообщества.

## 7.2. Современное состояние машинного перевода

Что же мы можем видеть сейчас, через почти пятьдесят лет с того момента, когда на машине IBM 701 русский текст «Величина угла определяется отношением дуги к радиусу» был переведен на английский язык без участия человека, чем состоялся так называемый Джорджтаунский эксперимент, начавший эпоху машинного перевода? Мы видим довольно пеструю и противоречивую картину.

С одной стороны, машинный перевод есть. В России (в Москве) можно сейчас купить приличную систему машинного перевода с русского языка на английский (и с английского на русский) под названием «Сократ» за 149 долларов, а также системы машинного перевода «Стилус», «Промпт» и др. Система МП фирмы «Бит» стоила 300 долларов. Системы МП в США стоят от 1000 долларов и качество перевода не лучше, чем дают наши системы. Эти затраты несовместимы с расходами на разработку систем машинного перевода «с нуля». Особенно дорого и сложно разработать эффективное программное обеспечение. Демонстрация работы систем показывает, что перевод получается удовлетворительный, даже тексты, принесенные потенциальным заказчиком и никак не учтенные при разработке и составлении словарей, иногда получают приличный перевод. Однако в большинстве случаев перевод требует постредактирования. Работать с современной системой МП переводчику, незнакомому или плохо знакомому с программированием и с работой на компьютере, довольно сложно. Системы, как правило, обладают защитой от перепи-

сывания, вносить изменения в алгоритмы и программы нельзя. Словари пополнять и изменять можно, однако часто лишь с помощью специальных программ, продающихся отдельно. Ввод текста — через сканер, с ручным исправлением искажений. Эта часть тоже продается отдельно. Системы допускают постредактирование.

Несмотря на невысокую стоимость, приемлемое качество и потребность в переводах, объем продаж, как кажется, невелик, и, насколько известно, не так много фирм пользуются машинным переводом. Такое положение можно объяснить следующим:

1) переводом в фирме занимается переводчик, он системным программистом не является, и свободное обращение с программой перевода, насколько бы дружественной к пользователю она ни была, для него дело довольно сложное. Прикомандировывать программиста — возрастают расходы. Редактировать плохой машинный перевод есть занятие трудоемкое и весьма неохотно выполняется редакторами и переводчиками, которым часто проще перевести текст заново, «с нуля», чем машинным образом выискивать и исправлять ошибки;
2) переход на другие тематики и другие типы текстов, что весьма распространено в переводческой деятельности и осуществляется человеком без особого труда, в машинном варианте всегда почти довольно затруднителен и может быть вообще невозможным при больших различиях в текстах.

Учитывая невысокое качество перевода и высокую стоимость обслуживания, другие неблагоприятные факторы нынешней информационной ситуации, применение современных действующих систем МП не является распространенным и их вклад в преодоление языковых барьеров (хотя точные цифры и отсутствуют), видимо, невелик.

Тем не менее, принимая во внимание общую ситуацию с переводами в мире, можно смело утверждать, что альтернативы машинному переводу нет. Какие соображения говорят в пользу такого утверждения?

## 7.3. О преодолении языковых барьеров

Современная языковая ситуация в мире характеризуется наличием большого количества естественных языков, на которых осуществляется мировая коммуникация во всех областях человеческого взаимодействия: наука, техника и технологии, политика, социальная практика, образование и т. д. Исследователи полагают, что число языков, на которых говорит человечество, исчисляется несколькими тысячами: точное число зависит от того, какое количество говорящих на том или ином языке как на родном мы принимаем в качестве значащей величины. Наибольшее количество говорящих в мире — на китайском языке, далее следуют Индия и другие страны с большим количеством населения. Однако количество печатной продукции не находится в пропорции с количеством говорящих на том или другом языке: так, по данным ЮНЕСКО, наибольшее количество публикаций в мире выходит на английском, далее следует русский язык, затем другие европейские языки (Марчук 1992). В любом случае число публикаций в мире велико, возрастает и осуществляется на большом количестве языков. Языковые барьеры, таким образом, являются главными на пути распространения научной, технической и всякой другой информации, на пути взаимодействия человеческого общества, социальных и экономических контактов.

Способов преодоления языковых барьеров можно насчитать три:

1) культивация и использование единого универсального языка, по крайней мере для основного состава коммуникантов;
2) изучение иностранных языков;
3) перевод.

Рассмотрим эти способы.

*Единый универсальный язык.* В этом способе возможны два варианта:

- **единый искусственный универсальный язык,**
- **один из естественных человеческих языков.**

Мечта о едином универсальном языке с незапамятных времен занимала умы ученых. Ньютон, Лейбниц, Декарт, а еще ранее — Беда Достопочтенный и другие выдающиеся умы человечества изучали и разрабатывали эту идею (Денисов, 1965). Несмотря на то, что Джонатан Свифт высмеял эту идею, придумав Лапуту и лапутян, она внесла значительный вклад в развитие представлений о логике в естественном языке, языковых универсалиях и в других аспектах лингвистической, философской и точных наук. К настоящему времени человечество имеет целый ряд искусственных языков, некоторые из них довольно известны и распространены, например эсперанто. Во многих странах есть клубы эсперантистов, выходят издания на этом языке. Однако руководители информационных служб, больше всего озабоченные преодолением языковых барьеров, отрицательно оценивают перспективы эсперанто и других искусственных языков в части результативности преодоления таких барьеров. Мало кто хочет изучать неродной язык, на котором не говорит ни один народ. Попытки создания новых, более удобных, по мнению авторов, искусственных языков, продолжаются, однако вряд ли с этим стоит связывать практические перспективы широкого применения универсальных искусственных языков.

Вариант второй — распространится один из естественных языков, на котором будет общаться человечество. В настоящее время мы имеем широкое распространение английского языка, на котором выходит наибольший объем публикаций в мире и который преобладает в международном общении. Станет ли английский язык единым универсальным мировым языком? Руководители информационных служб отрицательно отвечают на этот вопрос. С ростом национального самосознания все новых государств, с развитием науки и техники в этих государствах, повышением их жизненного уровня, возрастает потребность общения на своем родном языке. Мы видим такую ситуацию в бывших колониях европейских стран, среди государств, обретших национальную независимость. Чем выше развитие страны, тем больше усилий она готова затратить на право говорить и общаться на своем языке. Опыт работы крупных международных сообществ, таких, как ООН, Евро-

пейский Союз и др., показывает, что многоязычие в таких организациях неизбежно. В каждом из них существует несколько официальных языков.

Таким образом, единый язык как средство преодоления языковых барьеров не может считаться универсальным и перспективным. При этом, по моему мнению, работы в направлении создания искусственных языков следует приветствовать, поскольку они дают интересный материал для интерлингвистики, которая занимается проблемами сопоставления разных языков и нахождения языковых универсалий (Марчук, 2004).

*Изучение иностранных языков.* В каждой стране мира, в каждой школе изучаются иностранные языки. По данным специальных обзоров, школьники учат как минимум один иностранный язык, как максимум — два или три. Однако что объединяет все школы всего мира — это слабое знание этих языков учащимися после окончания школы. Единственными исключениями бывают школы многоязычных стран вроде Швейцарии или Люксембурга или семьи с разноязычными родителями и т. п. Лишь в редких случаях школьное изучение иностранных языков может быть эффективным способом преодоления языковых барьеров.

Изучение иностранных языков продолжается и на последующих ступенях образования и подготовки специалистов. В вузах практически всех стран производится языковая подготовка, есть и отдельные образовательные компоненты, учащие иностранному языку. Тем не менее, как следует из соответствующих обзоров и анализов ситуации, лишь немногие специалисты из общей их массы владеют иностранными языками в достаточной степени для самостоятельного специализированного общения. Поэтому такое знание как универсальное средство также не может быть обозначено.

Что же остается для преодоления языковых барьеров? Остается третий способ — *перевод.*

*Перевод* является видом информационной деятельности, потребность в которой никогда не сокращается (только при условии вымирания человечества не нужен будет перевод на

разные человеческие языки). Если рассматривать перевод как товар, то перепроизводства этого товара никогда не будет, всегда остается спрос для этого товара. Исследования рынка переводов показали, что этот рынок увеличивается примерно на 15 % в год, при этом около 30 % рынка удовлетворяется за счет переводов, выполняемых незарегистрированным образом (членами семьи и т. п.). Все большее количество людей занято в переводческой деятельности. Бюджет Европейского экономического сообщества в административной части на одну треть состоит из расходов на перевод: каждый документ растущего документооборота этого многонационального объединения должен быть официально переведен на все другие языки сообщества (Better Translation, 1983). Такое же положение и в других многонациональных организациях, число и функции которых также увеличиваются. Переводческая деятельность в ООН, как хорошо известно, требует больших расходов и организационных усилий.

Основная масса переводов — это не художественный или драматический перевод, как часто представляют себе люди, не знающие истинного положения вещей. Подавляющая масса переводов — перевод научно-технический, научный, коммерческий, деловой, рекламный, юридический, политический. Отсюда необходимость оперативного производства соответствующих терминологических словарей. Для переводов такого рода необходимым условием является своевременность и быстрое их выполнение: если перевод какой-то научной статьи задержится на несколько лет, вся ценность научного изобретения или теории может быть совершенно утрачена. То же относится и к патентам, деловым материалам и многим другим видам перевода. По качеству перевод может быть достаточно детально дифференцирован: возможен информативный перевод, целью которого является общее знакомство с содержанием текста на иностранном языке, точный перевод, нужный для детального знакомства с содержанием, сверхточный и юридически удостоверенный перевод для официальных документов и законов. Варьируемость качества позволяет в той или иной мере использовать механические и компьютерные возможности для облегчения и ускорения процесса перевода и контроля его качества.

Однако увеличение объемов переводов, которое требуется для обеспечения многоязычной коммуникации в развивающемся и информатизирующемся обществе, вряд ли можно себе представить как достигаемое за счет увеличения количества людей-специалистов, занятых переводами и редактированием. Это было бы все равно как, согласно анекдоту, представлять запуск космического корабля посредством большой рогатки, натягиваемой огромным количеством людей, или, ближе к реальности, строительства пирамиды большим количеством технически несовременно вооруженных рабов. Естественно полагать, что все большая часть труда по переводу будет возлагаться на все более совершенные компьютеры. Чтобы это действительно произошло, нужно продолжать занятия машинным переводом как в теории, так и на практике. Этим и объясняется актуальность машинного перевода в настоящее время.

## 7.4. Основные проблемы современного машинного перевода

Представляется, что оптимальное решение проблемы машинного перевода, точнее, наиболее актуальное приложение усилий к решению этой проблемы, может быть достигнуто только в том случае, если мы определим приоритеты научных исследований в этой области.

Два основных направления возникают в этой связи:

- попытка сделать так, чтобы переводился смысл безотносительно к форме языкового высказывания,
- осуществление машинного перевода на уровне переводных соответствий, когда передаются в равной степени и форма, и содержание.

Казалось бы, эти два направления не противоречат, а как бы дополняют друг друга. Однако на самом деле в истории машинного перевода они постоянно находились в противоборстве, что в значительной мере мешало эффективному развитию исследований и разработок.

Первое направление можно обозначить как «смысл — текст» или «текст — смысл», а более точно, применительно к собственно переводу, «текст — смысл — текст». Исследования

в этом направлении имели целью разработать универсальный смысловой язык-посредник, а когда стала ясна недостижимость этой цели (довольно быстро, впрочем), стремились создать отдельный семантический компонент систем, чтобы он контролировал и переводил семантику как высший уровень языковой структуры в переводе. Это стремление также практически провалилось, ни одной действующей системы с семантическим компонентом мы до сих пор не имеем нигде в мире, хотя работы велись широким фронтом во многих странах, в том числе и в СССР.

Второе направление можно назвать моделированием перевода на уровне переводных соответствий. Исторически оно возникло из начальных воззрений инициаторов машинного перевода относительно языка как некоторой разновидности кода (Марчук, 1983). Кроме того, в машинном переводе решающая роль отводилась словарю. Сторонники передачи смысла иногда именовали данный подход «методом грубой силы» (brute force approach). Однако игнорирование этого подхода приводило вообще к отказу от машинного перевода. Что значит «передать смысл» в переводе? Смысл предложения, высказывания, абзаца и целого текста можно без особых трудностей выразить либо одним словом, либо парой слов, либо цифрой, кодирующей данный «смысл» в какой-то знаковой системе (например, информационного поиска). Но ведь в переводе должен быть передан не такой «смысл», а смысл на уровне и формы, и содержания. Кроме того, вообще понятие «смысл» достаточно неопределенно (см. выше раздел о семантическом анализе). В чем разница между содержанием и смыслом? Переводные соответствия уточняют этот вопрос. Поэтому отказ от учета именно переводных соответствий и означал отказ вообще от перевода.

Тем не менее долгое время и ценой многих усилий исследования велись в этом тупиковом направлении. Тупиковость его стала особенно очевидной именно сейчас.

В отличие от этого моделирование перевода на уровне переводных соответствий и разные модификации этого подхода (иконический перевод и т. п.) дали реально работающие системы машинного перевода и позволили осмыслить как теорию

МП, так и практику, намечая эффективные пути совершенствования (см. также: Пиотровский, 1999; Зубов и др., 2004).

### 7.4.1. Проблемы современного машинного перевода

Основные проблемы современного машинного перевода удобно рассмотреть на материале проблемного доклада К. Буатэ, известного французского специалиста по МП. К. Буатэ относится к числу сторонников использования смысла в машинном переводе в виде языка-посредника. Одновременно он в течение долгого времени занимался созданием систем МП, которые бы могли реально переводить тексты. В своем докладе в Пенанге на международной конференции по актуальным проблемам вычислительной лингвистики в Малайзии (Boitet, 1991) он сформулировал двенадцать проблем современного машинного перевода. Буатэ выделяет три вида современного машинного перевода: 1) «информативный» машинный перевод. Это грубый машинный перевод, пословный, достаточный для поверхностного знакомства с содержанием текста на незнакомом языке; 2) «профессиональный» МП. Качество перевода сравнимо с качеством «человеческого» перевода и при небольшом редактировании является полностью удовлетворительным. Выгоден для больших объемов текста (свыше 10 000 страниц в год) и для однородных текстов; 3) «персональный» МП. Авторы подлежащих переводу текстов заранее избавляют их от неоднозначностей и работают в режиме диалога с компьютером. Возникает возможность существенно улучшить качество перевода.

Двенадцать проблем современного машинного перевода Буатэ делит на четыре класса: концептуальные проблемы, проблемы архитектуры, инженерные проблемы и технические проблемы. Имеет ли все это отношение к лингвистике? Прежде чем отвечать на этот вопрос, удостоверимся, что согласно принципам «нового структурализма» (да и старого здравого смысла) к объектам лингвистики принадлежат объекты, изучаемые лингвистом-практиком, которым лингвист-теоретик может дать теоретическую оценку. Эти объекты не могут ус-

танавливаться независимо от лингвистической практики (Prospects, 1992). Концептуальные проблемы формулируются К. Буатэ следующим образом:

- каким образом представлять и разрешать неоднозначности всякого рода? Каждой единице перевода сопоставляется, как правило, несколько возможных абстрактных представлений. Надо не только выбрать нужное в данном тексте, но и предвидеть появление других неоднозначностей в результате данного выбора. Для решения этого вопроса целесообразно ввести «неоднозначное программирование», предусматривающее разные возможности разрешения неоднозначностей;
- переводные соответствия. Как обеспечить точное соответствие между абстрактной структурой и текстом? Конкретное высказывание всегда отличается от его абстрактной модели. Концептуальной проблемой является также использование формализмов. Каким образом создать эффективные средства анализа из чистых формализмов, добавляя к ним процедурные или эвристические знания? Ни одна из действующих систем МП не использует только декларативные знания.

Если из этого текста вычленить «неоднозначное программирование», смысл которого плохо понятен лингвисту и которое является, на мой взгляд, не более чем некоторым спасательным кругом, помогающим математику и программисту, которым является Буатэ, не утонуть в проблеме, то остается следующее:

1) каждая единица перевода может иметь несколько абстрактных представлений;
2) надо еще предвидеть появление других неоднозначностей при движении далее по тексту от найденной, а также учесть в максимальной степени все другие возможные неоднозначности, могущие возникнуть в результате принятого решения;
3) у текста (не только у единицы перевода) может быть несколько абстрактных структур и неизвестно, как найти нужную;
4) чистый формализм не может служить основой анализа;

5) к формализму должны быть присоединены декларативные или процедурные лингвистические знания;
6) способ присоединения таких знаний неясен.

### 7.4.2. О переводе смысла

Пессимист, знакомый с историей машинного перевода, может сказать, что эти проблемы были ясны по крайней мере тридцать лет тому назад и что мы пришли опять к тому, от чего, вроде бы, должны были продвинуться. Однако оптимист, также знакомый с историей машинного перевода, заметит, что эти изложенные здесь проблемы сформулированы исследователем, всю свою деятельность посвятившим машинному переводу с помощью языка-посредника. В рамках этого подхода предполагалось перевести входной текст сначала на язык-посредник, а затем с него на любой другой естественный язык — выходной. При таком осуществлении МП все названные выше проблемы оказались бы за скобками — причем тут какие-то неоднозначности, если переводится смысл, а само понятие смысла исключает неоднозначности на всех предшествующих уровнях. Если мы знаем смысл высказывания, то все неоднозначности уже разрешены. В чем же тут дело?

Рассмотрим сначала концепцию языка-посредника для МП. Известно, что под таким языком понималось многое, от языка семоглифов до пучков грамматических соответствий, которые (как пучки, так и семоглифы) никто, включая авторов проектов, не знал, как строить и как практически применять к переводу (Мельчук и др., 1967). Дело, однако, было не в этом, а в обосновании концепции «машинный перевод без перевода, без машин, без алгоритмов». Переоценка возможностей компьютера дорого обошлась науке и значительно задержала исследования и разработки практически ориентированных систем. Многие подобные абстракции давно и благополучно забыты. Рассматривая вышеизложенные постулаты Буатэ, важно учесть прежде всего то, что они сделаны единственным последовательным сторонником идеи машинного перевода через язык-посредник, в функции которого предполагалось использование некоторого смыслового языка.

Трудности перехода на язык смысла от нормального естественного языка чрезвычайно велики. Десятилетия работ в области МП показали, что обращение к смыслу переводимого, такое вроде бы простое и естественное для человека-переводчика, в машинном моделировании чрезвычайно сложно. Начать с того, что понятие «смысл» трудно вразумительно определить. В. А. Звегинцев убедительно показал, что предложение, взятое вне контекста, имеет не смысл, а «псевдосмысл», а разные перефразировки одного и того же высказывания имеют и разный смысл. Он приводит перефразировки простого предложения: «Охотник изо всей силы ударил волка ногой» — «Волк получил удар ногой от охотника. Волк был повергнут ударом ноги существа мужского пола. Внезапный удар ногой сразил волка» и т. д. Содержание действий, соотношение субъекта и объекта вроде бы одно и то же, но одинаков ли языковый смысл вариантов высказывания? Без особых мудрствований видно, что нет. (Звегинцев, 1976). А что такое смысл текста? Это сумма смыслов отдельных высказываний? Если сумма, то какая — арифметическая или некоторая интегральная, зависимая от ситуативного и коммуникативного контекстов или нет? — и т. д. Список вопросов, на которые нет вразумительных и конструктивных с лингвистической точки зрения ответов, можно продолжить. Характеристикой положения является тот факт, что до сего времени нет ни одной действующей системы машинного перевода, которая бы в явном виде использовала понятие смысла и/или семантического компонента в объеме, большем чем для двух-трех примеров.

Что все это означает с точки зрения теории МП? Это прежде всего означает, что победила теория и концепция «текст — текст», а не «текст — смысл — текст». Действующие системы МП реализуют так или иначе принцип перевода с помощью переводных соответствий, без всяких семантических заглублений, больших, чем необходимо для устранения чисто переводных многозначностей.

Однако идея логического анализа естественного языка, исходящая из предпосылки, что любое высказывание строится по логическим законам и его можно разложить на некоторые исходные составляющие, с которыми далее можно работать по

законам логики, продолжает существовать в умах математиков и логиков. Можно предположить, что логический анализ естественного языка если и не даст в ближайшем будущем действующих систем машинного перевода, позволит понять и усовершенствовать структуру и принципы построения естественно-языкового интерфейса для общения с компьютером. (См. в этом аспекте работы: Шуклин, 2002; Петров и др., 1993).

### 7.4.3. Другие проблемы

Из этого следует, что теория машинного перевода должна развиваться в сторону дальнейшего осмысления эквивалентных, вариантных и трансформационных соответствий в рамках данной языковой пары с последующим переходом (когда нужен множественный перевод) к другой языковой паре. Не случайно Буатэ особо выделяет переводные соответствия, о которых он ранее не говорил. Думается, что теория должна выявить природу возникновения, характер, распространенность, возможности формализации таких соответствий, структуру и объем процедурных знаний, приписываемых каждому типу соответствий.

Рассмотрим теперь проблемы «архитектуры» в концепции Буатэ. Первой из таких проблем является соотношение между вариантами и настройкой. Вопрос ставится так: как конструировать и использовать грамматические формализмы, которые позволили бы описать варианты (микроязыки или подъязыки) в рамках единого языка, приспосабливая конкретную систему МП для данного естественного языка к конкретному подъязыку посредством настройки?

Здесь прежде всего выделяется лексический аспект. Невозможно содержать огромные базы лексических данных, которые были бы пригодны для каждого подъязыка в рамках данного языка. Совершенно также не исследован грамматический аспект (имеет ли каждый подъязык свою грамматику или же пользуется единой грамматикой языка?). Буатэ для решения этой проблемы предлагает объединение статистических и содержательных подходов. Так, статистика грамматических конструкций в рамках данного подъязыка может быть получена из

текстов. Содержательный анализ позволит выявить наиболее актуальные типы конструкций. Тогда бы подъязык мог быть определен как совокупность конечного инвентаря лексики и набора грамматических правил. Вопрос этот также ненов. Практика действующих систем машинного перевода пока еще не дала ответа на вопрос о том, как лучше организовать словари и грамматику в условиях разных подъязыков одного и того же языка. Можно лишь согласиться с Буатэ в том, что лингвистическая теория также еще не предложила каких-либо решений этой проблемы. Когда решение будет достигнуто, то еще предстоит научить пользователя выбирать тот или иной подъязык.

*Обучение на основе текстов.* Возможно ли создать системы машинного перевода, которые будут реально учиться на основе существующих параллельных текстов и на корректировке ошибок, отмеченных профессиональными редакторами?

Вопрос об обучении систем МП на основе параллельных текстов также ненов, однако реально мало что достигнуто на этом пути. Представляется интересным повторное обращение к статистическим методам в построении систем МП, в том числе и с интерлингвой, см., например последние японские работы, доложенные на Московской международной конференции по квантитативной лингвистике в 1994 году (Tsutsumi 1994). Буатэ также положительно оценивает роль параллельных текстов как при составлении словаря, так и грамматики. Остается лишь напомнить, что параллельные тексты служили основой и были главными критериями при разработке словарей и алгоритмов МП системы АМПАР еще в шестидесятых годах (Марчук, 1970; Моторин, 1970).

*Степень понимания.* Как воспроизвести «нечеткое» понимание человека-переводчика? Как воспроизвести высказывания частично в content-bound way, а частично в structure-bound way? Цитируем Буатэ. «Основанный на знаниях МП возможен в некоторых контекстах, где можно положиться на полное знание и полное представление некоторой предметной области или задачи. Например, можно переводить с помощью МП инструкции по уходу за ядерными реакторами. Но если пытаться

распространить МП на все области, то это невозможно». Нетрудно видеть, что это означает отказ от передачи смысла. Ведь в ограниченной предметной области, например в инструкциях, весь перевод реально осуществляется в рамках переводных соответствий и никакого обращения к смыслу (content) вообще не требуется.

*Связывание архитектуры с применениями.* Как увязывать архитектуру с типами применения? Здесь и далее излагаем концепцию К. Буатэ. Помимо своей существенной теоретической значимости, МП представляет собой инженерную проблему. Предложены многие архитектуры для разных применений (трансфер-интерлингва, пакетный-интерактивный режим и т. п.), не говоря о разнообразии лингвистических теорий и компьютерных инструментов. Но ни одна конкретная архитектура не годится для всех применений сразу.

Разработчики МП имеют свои идеи относительно адекватности архитектур и их применения. Проблема здесь заключается в согласовании действий разработчиков и заказчиков. Заказчик часто имеет свои собственные предпочтения, основанные на интуиции. Его нельзя переубедить в том, что такой-то путь неправилен, не пытаясь его реализовать. Но потом он теряет вместе с большими деньгами и весь интерес к МП.

Поэтому важная проблема для МП — дать четкую картину того, что можно ожидать и чего нельзя ожидать от МП.

*Измерение прогресса в МП.* Как измерить прогресс в МП и вклад исследований по МП в фундаментальные представления о характере языка и его использовании?

Наподобие использования термоядерной реакции, МП является мечтой нашего века. В противоположность термоядерной реакции, МП уже дает результаты. Однако финансируется МП далеко не так, как термоядерные исследования. Почему? Возможно, одной из причин является то, что отсутствует чисто научный элемент, а именно измерение прогресса. МП и связанные с ним исследования, видимо, не вносят какой-либо измеримый, фундаментальный вклад в науку.

Даже меньшую цель, а именно сравнение систем МП, трудно достигнуть. В контексте информативного МП может

быть организована конкуренция между системами, работающими в одинаковых языковых парах и в тех же тематических областях. Однако это пока что сделано, как сообщает Буатэ, только в отношении японско-английского МП.

В том, что касается профессионального МП, это почти невозможно, поскольку система МП может показать свою ценность только после настройки на конкретную область (по терминологии) и стиль (по грамматике). Такая специализация весьма дорога. Даже если подобное сравнение будет организовано на справедливой основе (финансирование нескольких исследовательских групп для специализации их систем), пока нет согласия относительно критериев, которые следует при этом применять. Суждения о качестве МП весьма субъективны. Измерения стоимости и эффективности за некоторый период времени более объективны, но требуют много времени для того, чтобы пользователь привык к системе. При этом важным критерием является легкость улучшения системы — пополнение словарей, повышение эффективности алгоритмов и пр.

С этими рассуждениями Буатэ нельзя не согласиться.

### 7.4.4. Современные системы машинного перевода

В настоящее время пользователю предоставляется множество систем машинного перевода, дающих достаточно грубый перевод, но тем не менее позволяющих преодолевать языковый барьер. П. Н. Хроменков в кандидатской диссертации исследовал наиболее известные системы МП, доступные пользователям, и показал, что с точки зрения теории практически все они так или иначе реализуют подход, который можно назвать «моделью переводных соответствий» в данной языковой паре (парах) (Хроменков, 2000). См. также: Марчук, 1997 относительно модели «текст — текст» в современных исследованиях и разработках. Довольно часто современные системы МП дают некоторый полупродукт, который далее редактируется человеком-редактором или переводчиком также с использованием компьютерных средств в виде автоматических словарей или так называемой «переводческой памяти» (translation memory). Подробно устройство современных систем машинного перево-

да на примере перевода с английского языка на русский, а также анализ немецкого языка с точки зрения машинного перевода и построения соответствующих алгоритмов рассмотрены в новейшем учебном пособии А. В. Зубова и И. И. Зубовой «Информационные технологии в лингвистике» (Зубов и др., 2004). В этом пособии подробно описывается процедура разработки как автоматических словарей, так и алгоритмов анализа и синтеза текста в машинном переводе.

В статье (Светова и др., 2002) описываются системы автоматизированного перевода со всякого рода вспомогательными средствами. Такое взаимодействие человека с информационными технологиями в переводе позволяет в первую очередь обеспечить быстрое выполнение переводов, что весьма существенно в информационную эпоху, и, кроме того, дает перевод достаточно высокого, во всяком случае приемлемого для заказчика качества. В сборнике Всероссийского центра переводов «Перевод: традиции и современные технологии» дается полезный обзор существующих систем и приемов автоматизированного перевода (см.: Светова и др., 2002).

Особо следует отметить вопросы машинного перевода с/на восточные языки. Эти языки в настоящее время выходят на мировую арену, и проблема их анализа и синтеза для МП, равно как и для других видов автоматической обработки, становится весьма актуальной. Перевод с участием этих языков интересен также и в плане сопоставительной лингвистики, поскольку различия в грамматическом строе, особенно в синтаксисе, между этими языками и европейскими значительны и требуют особых решений в МП. С другой стороны, формальное изучение различий и инвентаризация способов их преодоления для перевода дают интересный материал для обобщений в рамках общей лингвистической теории и позволяют заново оценить известные лингвистические позиции и проблемы, такие, например, как части речи (см.: Кривоносов, 2001). Вопросы машинного перевода с восточных языков на европейские и наоборот рассмотрены в работе (Марчук, 2002).

## 7.5. Машинный перевод как центральная проблема искусственного интеллекта

### 7.5.1. Место машинного перевода в ряду интеллектуальных задач

**Системы машинного перевода текстов с одного естественного языка на другой осуществляют сложные преобразования на всех языковых уровнях двух и более языков, ставя целью передачу смысловой информации текста одного языка и создания эквивалентного ему по форме и содержанию текста на другом, выходном, языке. В данном разделе мы хотим обратить внимание на место МП в ряду других задач искусственного интеллекта.**

Мысль о том, что машинный перевод является центральной проблемой искусственного интеллекта, принадлежит академику Юрию Владимировичу Рождественскому, много занимавшемуся вопросами новой фактуры речи — современными способами работы с текстами в новых условиях массовой коммуникации и культуры. Можно спорить о том, действительно ли именно машинный перевод является центральной проблемой моделирования человеческого интеллекта — искусственного интеллекта, хотя многие ученые вообще скептически относятся к формуле «искусственный интеллект». Тем не менее, никто не будет спорить о том, что перевод с одного естественного языка на другой действительно является сложнейшей проблемой, затрагивающей практически все аспекты человеческого мышления и способов языкового выражения мысли.

В связи с этим следует отметить новое направление прикладных лингвистических исследований — изучение естественного интеллекта. Н. К. Рябцева пишет: «... В результате естественный интеллект — *самый сложный и неисчерпаемый объект изучения* — стал, в определенном смысле, "фоновым", как бы исчерпавшим свои эпистемические возможности термином, а не "выделенным" и продуктивным понятием. Соответственно, "за кадром" остается масса важнейших вопросов: почему всякое обучение, строго говоря, не так эффективно, как хотелось бы? Почему автоматизированные системы обработки

информации на естественном языке, несмотря на гигантские усилия своих разработчиков, мягко говоря, далеки от совершенства? Как соотносятся такие "ментальные" явления, как сознание, подсознание, разум, рассудок, рассудительность, ум, умонастроение, познание, мышление, память, знания, взгляды, а также смекалка и т. п. и интеллект?» (Рябцева, 2004, с. 71). Язык и интеллект человека неразрывно связаны и потому представляют собой важнейший источник сведений друг о друге, говорит далее автор.

Целью машинного перевода, как и всякого перевода, является передача смысла. Для такой передачи необходима экспликация смысла, в том числе и средствами языка-цели. Описание смысла сообщения в терминах смыслового языка, например искусственного, есть тоже перевод. Между тем такое описание и извлечение смысла, эквивалентное распознаванию, пониманию, есть не что иное, как центральная задача моделирования мышления. Поэтому система машинного перевода, моделирующая одну из центральных задач познания мышления, есть одновременно система искусственного интеллекта.

Представляет интерес в этой связи концепция академика Ю. В. Рождественского (Рождественский и др., 1988; Рождественский, 2003). Он различает элементарные, базовые системы искусственного интеллекта, и сложные системы. Различие между этими системами состоит в назначении, характере операций и в информационном продукте. Сложные системы включают: а) моделирующие; б) экспертные; в) обучающие; г) игровые. Эти три класса систем основаны уже не на составлении или прочтении текста по буквам, как простые системы, а на семантике текстов. Для операций над семантикой текстов необходимо сопоставление текстов и их частей. Такое сопоставление возможно лишь при определенном отношении к содержанию текстов.

Переводное значение отличается от значения, выделяемого как понятийное или образно-понятийное. Переводное значение не имеет характера называния примера, установления имени. Переводное значение не возникает в процессе таких операций, как филологическое истолкование текста, интерпретация стиля источника или создание стиля перевода. По мнению Ю. В. Рож-

дественского, переводное значение существует только в переводе и представляет собой особую семантическую аспектацию слова и текста.

Многие системы автоматизированного перевода присоединяют к сопоставительной переводной системе систему истолкования текста как систему моделирования содержания текста. Они содержат также словарь для справок редактора перевода. Таким образом системы автоматизированного перевода соединяются с моделирующими и экспертными системами.

Для своего функционирования системы МП нуждаются в записи, редактуре, устном вводе текста, устном управлении, анализе и синтезе текста, т. е. система автоматизированного перевода основывается на базовых системах искусственного интеллекта. Это значит, что система МП находится в центре систем искусственного интеллекта. Являясь частью информационной деятельности, система МП порождает тексты информатики, может быть использована для обучения, соединяется с экспертными и другими системами искусственного интеллекта. Схематически такое положение может быть изображено следующим образом:

| | | | |
|---|---|---|---|
| Запись, редактура | Синтез, анализ текстов | Документы | Научная и техническая литература |
| Системы графического синтеза | АВТОМАТИЗИРОВАННЫЙ ПЕРЕВОД | | Массовая информация |
| | Тексты информатики | | |
| Обучающие системы | | | Моделирующие системы |
| | Экспертные системы | | |

*Рис. 19.* **Система автоматизированного перевода в ряду других систем искусственного интеллекта**

Лингвисты и филологи, хорошо знающие лингвистические, текстовые трудности понимания, интерпретации, перевода, анализа и синтеза естественно-языковых текстов, ставят обычно системы машинного и автоматизированного перевода в центр систем искусственного интеллекта. Дело здесь в том, что специалисты такого профиля хорошо представляют себе,

какие последствия может иметь неправильно понятое слово, не воспринятая в данной коммуникативной ситуации мысль, неверный синтаксический анализ, ошибочный синтез. Интересно отметить, что специалисты другого профиля в искусственном интеллекте считают проблему перевода каким-то побочным следствием кодирования или более широкой задачи общения с ЭВМ по тому или другому поводу. Два подхода, иллюстрирующие такую точку зрения, целесообразно рассмотреть здесь.

Например, А. Н. Заслонко и Г. К. Хахалин считают машинный перевод частным случаем лингвистического транслятора, рассуждая следующим образом (Заслонко и др., 1989):

Основной задачей машинного перевода является анализ и синтез текста на естественном языке (ЕЯ), т. е. задача лингвистической трансляции. Особенностью описываемого метода трансляции является наличие модели проблемной области, к которой относятся тексты на ЕЯ. ЕЯ-текст переводится в структуры на языке этой модели (М-язык) при анализе, а затем эти структуры переводятся на ЕЯ-тексты при синтезе. В качестве метода трансляции используется метод контекстного фрагментирования. Сам лингвистический транслятор представляет собой трехуровневую систему: лингвистический процессор, включающий анализатор и синтезатор, лингвистическую модель, включающую знания о грамматике и семантике, и ассоциированные процедуры, осуществляющие взаимосвязь декларативной и процедурной частей лингвистического транслятора. Такая структура транслятора выбрана с учетом возможностей его настройки на различные естественные языки, разные системы классификации лингвистических отношений и пр.

Лингвистическая модель разделяется на три составляющих: грамматическая модель, интерпретационная модель и модель проблемной среды. Для анализатора и синтезатора, работающих в режиме двуязычного перевода, первая и вторая составляющие различны, а третья является общей. Если лингвистическая модель для анализатора и синтезатора одна и та же, то транслятор работает в режиме перефразировки, что важно для его настройки и отладки.

Базовый компонент анализатора осуществляет декомпозицию простых полных предложений естественного языка на

фрагменты посредством контекстов грамматической модели (синтаксических правил, записанных в виде сети связных структур), интерпретирует эти фрагменты во фрагменты языка модели и компонует фрагменты этого языка в структуру, описывающую проблемную ситуацию, которая представлена текстом на ЕЯ.

Расширенный компонент анализатора предназначен для анализа сложных, осложненных, эллиптических и анафорических ЕЯ-предложений. С помощью этого же метода — с использованием контекстов другого уровня — сложные предложения разбиваются на простые фразы, и строится структура сложного предложения, которая композиционно определяется взаимосвязями фраз в предложении. Каждая фраза обрабатывается базовым компонентом, а результаты перевода на язык модели компонуются с учетом связей между фразами и на основе семантических контекстов. Для восстановления эллипсисов и разрешения анафор используются дискурс и механизмы сопоставления, позволяющие найти недостающие структурные элементы для эллипсисов или антецедент для ссылки.

Каждое естественно-языковое предложение, прошедшее через анализатор, представляется структурой на языке модели. Синтезатор по этим структурам строит предложение на другом ЕЯ, либо с учетом структуры предложения анализируемого языка, либо без учета. В последнем случае осуществляется пересказ, т. е. перевод по смыслу, соответствующий входному тексту, но по структурному выражению не совпадающий с оригиналом. В основе синтеза также лежит метод контекстного фрагментирования.

Разработанная версия лингвистического транслятора, под названием АДАЛИТ, предназначена для общения на проблемно-ориентированном ЕЯ конечных пользователей с интеллектуальными вычислительными системами, но может быть использована, по мнению авторов, и как самостоятельная система в рамках системы МП. Проверка лингвистического транслятора осуществлялась на решении нескольких задач: доступ к БД, кодирование медицинских диагнозов и пр., на нескольких проблемных областях (хранение кадровой информации, медицина, геометрические объекты и пр.). Возможность ис-

пользования АДАЛИТа для машинного перевода иллюстрируется проведенными экспериментами, в которых на вход анализатора поступали ЕЯ-предложения на английском и эстонском языках, а на выходе синтезатора выдавались предложения на русском языке. Версия транслятора реализована на СП ЛИСП: объем порядка 2 тысячи строк; среднее время анализа — 3 секунды на слово, синтеза — 1 секунда на слово на ЭВМ ЕС 1055 при объеме словарей порядка 500 слов естественного языка.

Расценивая это предположение, видим, что оно исходит из слишком оптимистической предпосылки о том, что словарь в 500 слов по качеству и лингвистическим коннотациям ничем не отличается от словаря в 5000 слов, а последний — то же, что и словарь в 50 000 слов. Такой взгляд кажется слишком упрощенным. Только в очень ограниченном подъязыке и узкой предметной области можно уверенно работать с таким словарем. При возрастании словаря за счет общеупотребительных, общетехнических и общенаучных слов неизбежно существенное расширение омонимии, лексической многозначности, диапазона синтаксических конструкций и разного вида неоднозначностей.

### 7.5.2. МП как «побочное следствие» преобразований

Другая концепция ускоренного достижения машинного перевода — вывод из некоторого обобщенного представления о сущности языкового знака. А. В. Танасиенко считает, что одна общая концепция языкового знака даст возможность осуществить машинный перевод как некоторое побочное следствие общих языковых преобразований (Танасиенко, 1989).

Назовем пару обобщенных объектов *а* и *б* модельной парой, если они имеют множество общих свойств и признаков, взаимно отражаемых ими. Каждый из этих объектов является моделью другого, а принцип, декларирующий относительность статуса модели, назовем принципом модельной относительности. Воспринимаемый субъектом материальный объект и его идеальный (перцептивный) образ в сознании субъекта есть не

что иное, как модельная пара, т. е. пара обобщенных объектов, находящихся в модельном отношении. Автор называет *автонимным знаком* материальный элемент такой модельной пары и автонимным образом — идеальный ее элемент. В термине «автоним» объединяются два эти понятия: автонимный знак — он материален, и «автонимный образ» — он идеален. Одним из простейших примеров естественно-языкового автонимного знака является фонема. Из такого определения автонимного знака следует, что любой доступный восприятию материальный объект можно считать автонимным знаком, если он хотя бы раз был представлен субъекту и сформировал перцептивный образ в его сознании, причем число таких представлений можно интерпретировать как некоторую характеристику его автонимной «знаковости». Материальные объекты, не удовлетворяющие этим условиям, например принципиально ненаблюдаемые или ни разу не встретившиеся субъекту, автонимными знаками не являются.

Свойства, присущие естественным языкам и делающие их способными выступать в качестве достаточно эффективного средства коммуникации, предполагают определенную компактность автонимной модели языка на высоком иерархическом уровне, что позволяет предположить возможность реализации человеко-машинной структуры стыковки автонимных структур высокого уровня, полученных автоматически для двух разных языков.

Автору этой концепции представляется принципиально возможной реализация системы, способной автоматически строить по потоку текстов адаптивную иерархию языковых автонимов, отражающую систематическую информацию, заключенную в тексте, и в этом смысле реализующую машинную процедуру понимания. Автоматическое наращивание иерархической автонимной структуры должно обеспечиваться за счет работы алгоритмов частотного анализа.

Вследствие концептуальной компактности естественного языка, при правильном отборе алгоритмов, количественный рост автонимов не должен прогрессировать геометрически, но, напротив, должен зафиксироваться по достижении определенного иерархического уровня.

Система автоматического перевода с одного естественного языка на другой может быть получена в результате соединения (человеком, владеющим обоими языками) двух автонимных структур, выращенных для данной пары языков автоматически. После соединения автонимных структур такая система должна обеспечивать автоматический перевод для двух конкретных языков как в прямом, так и в обратном порядке. Перевод должен поддерживаться только за счет работы алгоритма, моделирующего распространение возбуждения в автонимной структуре под воздействием входного текста.

За счет реализации описанного механизма система, по мнению автора данной концепции, может обойтись без таких атрибутов систем машинного перевода традиционного типа, как жесткие алгоритмы синтаксического и морфологического анализа, а также машинные словари фиксированных словоформ, подготавливаемые человеком вручную.

### 7.5.3. Возможно ли такое решение?

Оценивая данное предложение с точки зрения практических аспектов машинного перевода, следует отметить, что пока что нет ни одной действующей системы МП, которая была бы построена только на основе чисто глобальных теорий, подобной этой. Дело в том, что автоматическое построение систем машинного перевода, которое заложено в подобной концепции и при котором учитываются некоторые общие свойства текстов и языковой действительности, обычно натыкаются на преграды, состоящие в том, что языковые структуры весьма индивидуальны и ориентированы в подавляющем большинстве случаев не на классы явлений, а на отдельные явления и языковые факты, такие, например, как отдельные слова, каждое из которых представляет собой самостоятельную проблему для анализа и синтеза и не укладывается в какие-то общие всеохватывающие схемы.

Таким образом, общеязыковые преобразования не являются достаточными для того, чтобы построить машинный перевод как некоторое частное следствие из общих теорем. Как правило, проекты такого рода выдвигаются либо специалиста-

ми по логике, либо программистами, не представляющими трудности конкретной общеязыковой задачи, каковой является перевод. Главная информация, подлежащая переводу, заключается не в общем, а в наборе индивидуальных языковых средств. А именно эти средства и игнорируются общими универсальными теориями преобразований. Не случайно важными моментами новой теоретической лингвистики провозглашены, как мы выше отмечали, основополагающие значения языковых частностей и практических решений.

## 7.6. Системы машинного перевода АМПАР и СПРИНТ

Система машинного перевода, служившая прототипом системы АМПАР, создавалась коллективом разработчиков под руководством Ю. А. Моторина в течение нескольких лет, начиная с 1956 года. Эти работы начались одновременно с исследованиями по машинному переводу в Институте точной механики и вычислительной техники Академии наук СССР, где ими руководили Д. Ю. Панов (со стороны вычислительной техники и программирования) и И. К. Бельская (лингвист). Создавался алгоритм перевода с английского языка на русский, результаты работы над которым в довольно полном виде были опубликованы позднее (Бельская 1969). В 1974 году, когда был образован Всесоюзный центр переводов научно-технической литературы и документации (ВЦП) Государственного комитета СССР по науке и технике и Академии наук СССР, а в нем — отдел машинного перевода, руководство которым было поручено автору этой книги, эта система была предоставлена в порядке технической помощи данному отделу, и началось ее программирование для тогдашних серийных ЭВМ общего назначения серии ЕС (единая система), для ЕС-1040 и ЕС-1065. Одновременно с программированием производилось совершенствование словарей и алгоритмов системы и настройка ее по тематике на английские тексты по вычислительной технике и программированию. Тогда же появилось название системы АМПАР, что означает «автоматизированный машинный перевод с английского на русский». Слово «авто-

матизированный» как бы подразумевает, что перевод требует редактирования человеком и не является полностью автоматическим в этом смысле.

Теперь можно видеть, насколько сильно была в то время скомпрометирована идея машинного перевода чрезмерным выпячиванием концепции «передачи смысла». «Перевод без перевода, без машин, без алгоритмов» утвердился в академической науке и подавлял здравомыслие в составлении словарей, описаний и алгоритмов. Декларировалось, что составление алгоритма — не дело рук лингвистов, пусть этим занимаются математики, а лингвисты должны лишь давать описание языковых фактов. Шапкозакидательские концепции, подобные этой и языку-посреднику в виде «пучков грамматических соответствий», надолго затормозили отечественные исследования, и только с образованием ВЦП возобладал практический подход, основанный на здравом смысле, хорошем знакомстве с возможностями вычислительной техники и программирования того времени и правильной постановке цели.

В настоящее время система машинного перевода АМПАР породила серию систем машинного перевода на персональном компьютере под названием СПРИНТ. Системы СПРИНТ в настоящее время проходят стадию совершенствования программного обеспечения.

## 7.7. Система машинного перевода САПФИР

### 7.7.1. Общая концепция

Система машинного перевода САПФИР разработана коллективом исследователей МГУ под руководством проф. Ю. Н. Марчука. Программная часть и концепция программного обеспечения создана на базе программного инструментального комплекса ЛИНГВИК (автор В. И. Шлейников).

Данная система была одной из первых систем МП, ориентированных на персональные компьютеры. Если практически все системы машинного перевода, начиная с 1956 года, строились с расчетом на использование на больших компьютерах, то

в конце восьмидесятых годов произошла переориентация этих систем на персональные компьютеры, которые к тому времени по объему памяти и быстродействию сравнялись в большими ЭВМ и могли предоставить пользователю гораздо большие удобства, чем большие машины. В связи с этим уже в координационных планах работ по машинному переводу в СССР появились первые работы, связанные с созданием систем МП для персональных компьютеров (Марчук, 1996).

Специализированный инструментальный лингвистический комплекс ЛИНГВИК представляет собой параметрически настраиваемую транслирующую систему, функционирование которой полностью определяется заданными в декларативной форме описаниями, отражающими все этапы выполнения процессов преобразования текста с заданного входного языка на заданный выходной язык.

### 7.7.2. Процесс трансляции

Выполнение процесса трансляции текста с заданного входного языка на выходной система ЛИНГВИК организует в соответствии со ставшей уже традиционной в построении трансляторов схемой, по которой процесс трансляции разделяется на три основных этапа:

- перевод исходного (входного) текста в некоторую промежуточную форму представления — цепочку лексем;
- построение на базе полученной цепочки лексем дерева связей элементов транслируемого текста (дерева вывода);
- генерация выходного текста — результата работы транслирующей системы.

Управление функционированием указанных этапов осуществляется при помощи определения четырех взаимосвязанных описаний.

Функционирование первого этапа задается описанием словаря разрабатываемой транслирующей системы, которое определяет множество распознаваемых системой слов — терминальных символов, характеризующие эти слова наборы признаков (атрибуты), а также зависимые от атрибутов возможные переводные эквиваленты для этих слов.

Функционирование второго этапа задается описаниями множеств правил, определяющих синтактико-семантическую структуру правильных фраз входного языка и связанных с ними правил вычисления атрибутов в процессе построения дерева связей.

Функционирование третьего этапа задается описанием множества взаимосвязанных с синтактико-семантическими правилами правил генерации выходного текста (правил семантического синтеза).

Самым сложным и трудоемким этапом в схеме является второй, эффективность реализации которого и определяет эффективность работы всей системы в целом. Это было учтено при создании системы ЛИНГВИК. В частности, был выбран метод, основанный на применении атрибутных трансляционных грамматик, который успешно применялся в создании ряда трансляторов и показал свою высокую эффективность. Однако входные языки таких трансляторов относились к классу искусственных алгоритмических языков, что облегчало задачу построения транслирующих систем. Здесь же, в задаче машинного перевода, требуется приспособить данный комплекс к переводу естественного языка.

Описание процесса трансляции состоит из следующих четырех взаимосвязанных компонентов:

1) описание словаря входного языка (лексический компонент описания входного языка);
2) описание синтактико-семантической структуры фраз входного языка (синтактико-семантический компонент описания входного языка);
3) описание правил вычисления атрибутов при проведении синтактико-семантического анализа;
4) описание правил синтеза (порождения) фраз выходного языка (компонент семантического синтеза выходного языка).

### 7.7.3. Содержание этапов трансляции

Рассмотрим кратко эти основные компоненты.

Лексический компонент описания входного языка представляет собой множество, состоящее из нескольких словарей,

каждый из которых имеет свой номер и соответствует отдельной смысловой подобласти рассматриваемого входного языка.

Словари с номерами 0 и 1 имеют особое назначение. В словаре с номером 1 хранятся слова, используемые как служебные в других словарях, словарь с номером 0 содержит слова, несущие одинаковую смысловую нагрузку во всех смысловых подобластях входного языка, в то время как слова в других словарях имеют семантическое значение только в пределах соответствующей смысловой подобласти входного языка и теряют его в других подобластях. Это, в частности, позволяет одни и те же слова, но включенные в разные словари, использовать с разной смысловой нагрузкой внутри соответствующих смысловых подобластей входного языка.

Внутри каждого словаря слова могут объединяться в группы семантической эквивалентности, т. е. все слова, входящие в такую группу, считаются семантически эквивалентными в пределах соответствующей данному слову смысловой подобласти. В качестве примера группы семантически эквивалентных слов при описании ограниченного множества слов английского языка можно привести группу, куда входят артикли.

Условным обозначением семантики включенного в конкретный словарь слова в общем случае считается само знаковое представление этого слова, а для групп семантической эквивалентности — знаковое представление первого включенного в эту группу слова. Однако для удобства организации процесса генерации выходного текста пользователю ЛИНГВИКа предоставляется возможность самому определить условное обозначение семантики включенных в словарь слов или групп семантически эквивалентных слов путем задания перед словом или перед первым элементом группы семантически эквивалентных слов своего собственного условного семантического значения, предваряя его особым символом. Понятие условного семантического значения слова является весьма важным для организации функционирования транслирующей системы, так как позволяет графическому написанию слова входного языка поставить в соответствие любую заданную информацию, которая может быть обработана или выделена в любой файл, в том числе и на вывод, специальными процедурами системы ЛИН-

ГВИК. В частности, при реализации на базе ЛИНГВИК системы машинного перевода с помощью задания условных семантических значений решается задача определения для слов входного языка основных переводных эквивалентов, обрабатываемых в зависимости от значений вычисленных в процессе анализа входного языка атрибутов.

Помимо основного семантического значения для каждого включенного в словарь слова можно задать набор дополнительных признаков — атрибутов морфологического или синтактико-семантического характера (например, признаки обусловленности, рода, числа и пр.). Эти признаки представляют собой дополнительную информацию о конкретном слове словаря, которая может быть использована как для организации выдачи различных переводных эквивалентов для одного и того же слова, так и для организации проведения более глубокого анализа смысловой нагрузки, которую данное слово несет в анализируемом входном тексте.

Набор признаков, которыми могут снабжаться включаемые в словарь слова, задается разработчиком транслирующей системы и описывается в словаре с номером 1. Для каждого признака (или группы признаков) разработчик выбирает удобное для дальнейшего использования имя и связывает с этим именем соответствующий код признака (или группу кодов признаков).

Новый подход к созданию алгоритма автоматического перевода предполагает и свой способ организации словаря. Особенности такого словаря определяются двумя основными факторами.

Первый из них связан с ответом на вопрос о том, какая необходимая информация должна отражаться в словаре.

В системе ЛИНГВИК каждому слову входного языка может быть поставлено в соответствие некоторое условное семантическое значение, которое можно трактовать как переводной эквивалент слова. Кроме того, слово входного языка может быть снабжено некоторым набором признаков — атрибутов. Термин «атрибут» в данном случае практически синонимичен термину «признак» и употребляется только для того, чтобы сохранить терминологию, которая используется в опи-

сании атрибутных трансляционных грамматик. Значения признаков слова хранятся в словаре и используются в процессе получения текста перевода.

### 7.7.4. Примеры возможных атрибутов и их значений

Существительные в русском языке могут иметь атрибуты рода, числа и падежа. Значениями этих атрибутов у рода будут мужской, женский и средний. У числа — единственное и множественное число, у падежа — именительный, родительный и пр. Кроме того, в словарь должна быть включена информация о различных формах слова, его окончаниях. Окончания или словоформы с приписанными к ним значениями атрибутов составляют парадигму слова. Таким образом, в словарь входит множество английских и русских слов (поскольку система в нынешнем виде предназначена для этих двух языков) вместе с окончаниями и/или словоформами, а также значениями атрибутов. Русские слова при этом записываются в качестве условных семантических значений английских слов и выступают в роли их переводных эквивалентов.

После определения состава информации в словаре следует ответить на другой вопрос, а именно: какой порядок организации словарной статьи наиболее удобен?

Один из способов организации словарной статьи следующий. Словарная статья разбивается на две части — английскую и русскую. Английская часть представляется в форме описания слов входного языка, а русская — в форме описания их условных семантических значений. Порядок организации и той и другой частей практически одинаков. Информация к слову имеет свои правила записи, они таковы.

1) Любое слово может иметь множество значений атрибутов. Например, глагол имеет атрибут времени, значениями которого в русском языке будут настоящее, прошедшее и будущее время. Эти значения включаются в словарь и могут быть использованы в процессе работы системы.
2) Любое из значений атрибутов может включать подмножества, состоящие из наборов значений других атрибутов.

Такие подмножества заключаются в квадратные скобки. Пример: глагол в настоящем времени имеет атрибуты единственное число и множественное число. В свою очередь эти атрибуты составляют подмножество атрибута настоящее время, а каждый элемент такого подмножества может включать в себя любое другое подмножество. Например, атрибут единственное число имеет подмножество, состоящее из атрибутов первого, второго и третьего лица. То же происходит и с атрибутом множественного числа.

3) Если набор атрибутов характеризует окончание или словоформу, то он приводится после соответствующих атрибутов.

Большинство слов русского и английского языков можно разделить на группы в зависимости от имеющихся у них парадигм. Слова с одинаковыми парадигмами будут относиться к одной группе, с разными — к разным. Такое деление позволяет упростить описание словаря, а именно: в словарной статье не будет выписываться каждый раз информация, касающаяся атрибутов окончаний и словоформ, а будет даваться только ссылка на метку, указывающая на соответствующую парадигму. Слова, ссылающиеся на одинаковую метку, имеют и одинаковую парадигму. Таким образом, словарь состоит из множества словарных статей, а также общего списка парадигм. Что касается слов с нерегулярными формами словоизменения, то их парадигма приводится непосредственно в словарной статье.

### 7.7.5. Описание синтактико-семантической структуры входного языка

Это описание базируется на понятии атрибутной трансляционной грамматики. Однако в системе ЛИНГВИК это понятие было существенно модифицировано и расширено, что позволило получить принципиально более широкие возможности для построения на базе ЛИНГВИК разнообразных транслирующих систем.

В принятой в системе ЛИНГВИК интерпретации атрибутная транслирующая грамматика состоит из правил следующих трех видов:

1. Правил контекстно-свободной грамматики, при помощи которых система строит дерево вывода для транслируемых предложений.
2. Правил вычислений и согласований атрибутов.
3. Правил генерации выходного текста — результат работы транслирующей системы.

Важная роль в атрибутной трансляционной грамматике отводится понятию атрибут. Атрибуты — это элементы некоторого множества, отражающие те или иные признаки обрабатываемых объектов: слов, правил и т. д.

Атрибуты тесно связаны с деревом вывода, которое строится системой для обрабатываемых предложений. Атрибуты бывают двух типов: унаследованные и синтезированные. Первые зависят от непосредственного предка соответствующего узла дерева вывода, вторые — от атрибутов непосредственных потомков узла. Унаследованные и синтезированные атрибуты получают значения, переданные от окружающих ветвей дерева вывода. Таким образом, унаследованные атрибуты содержат информацию, переданную вниз от корня дерева вывода, а синтезированные — ту, которая передается вверх к корню дерева вывода.

Рассмотрим более подробно атрибутную трансляционную грамматику, на базе которой построен алгоритм переводных преобразований.

Атрибутная трансляционная грамматика — это грамматика, строящаяся на множестве терминальных и нетерминальных символов, а также на множестве операционных символов, и обладающая следующими свойствами:

1) каждому терминальному или нетерминальному символу может соответствовать некоторый набор атрибутов из определенного в системе множества атрибутов;
2) все атрибуты могут быть либо унаследованными, либо синтезированными;
3) значения унаследованных и синтезированных атрибутов вычисляются в процессе проведения синтактико-семантического анализа транслируемого текста по специально заданным правилам.

Грамматика называется корректной тогда и только тогда, когда для любой выходной цепочки можно однозначно построить дерево вывода и вычислить все значения всех атрибутов каждого нетерминального символа.

Правила первого вида, т. е. контекстно свободные правила, предназначены для описания предложений входного языка (в данном случае английского). Они имеют вид X → Y, где Y может состоять из нескольких компонентов, каждый из которых может замещать какой-либо другой в зависимости от правил подстановок.

В правила могут входить терминальные и нетерминальные символы. Примером терминального символа может служить символ BE, введенный для обозначения всех форм глагола to be (am, is, are, was, etc.), в том числе и инфинитива. Нетерминальные символы не являются словами входного языка. Они могут обозначать типичную конструкцию или группу зависимых слов, а также иногда служат названием целого класса слов, например класс определителей, куда входят артикли, указательные (this, these, that, those) и притяжательные местоимения.

Структура предложения находится путем многократного применения правил подстановки.

Мы говорим, что последовательность правил R1, R2,... Rn есть вывод структуры предложения S, если R1 есть исходное правило, а все остальные, т. е. R2, R3,... Rn правила раскручивают R1 до терминальных символов, образующих цепочку S, состоящую из слов входного языка.

Под предложением в данной работе понимается упорядоченное множество слов, отделенных друг от друга пробелами и оканчивающихся специально выбранным символом конца — точкой. При этом словами являются не только слова в обычном понимании, но и распознаваемые знаки препинания (запятая, двоеточие и пр.).

Для того чтобы описать любое предложение входного языка средствами атрибутной грамматики, прежде всего необходимо выделить типичные составляющие предложения и описать их соответствующими правилами. В общем случае это становится возможным благодаря структурному подходу, ко-

гда язык представляется совокупностью составных частей, связанных определенными отношениями. Будем считать, что основной единицей, выражающей смысл, является предложение. По структуре смысл предложения можно трактовать как систему понятий, отражаемых словами. Эти слова-понятия могут объединяться в словосочетания по правилам формирования отношений между понятиями, образуя новые понятия. Отсюда получается следующая схема структурных единиц языка:

< предложение >
< словосочетание >
< слово >

В данной грамматике составляющая является промежуточным звеном между словом и предложением. Это отличает данный подход от тех систем машинного перевода, в которых за основу берется только синтагма — сочетание двух слов, связанных определенным отношением, и в которых анализ предложения производится путем выявления всех входящих в него синтагм и построения на этой основе дерева зависимостей. Принятый в системе ЛИНГВИК подход позволяет решать те же задачи, но, кроме того, открывает возможность для решения более сложных задач.

Прежде чем разрабатывать алгоритм анализа, необходимо провести тщательное исследование и описание типичных структур входного языка. Для удобства можно принять, что основным элементом описания предложения атрибутной грамматикой является составляющая. В этом случае главной лингвистической задачей является определение типичных структур входного языка. Но поскольку любая грамматика, адекватно описывающая язык, должна иметь возможность охарактеризовать практически любое правильно построенное предложение, то возникает вопрос о том, как быть с единичными явлениями, например фразеологическими оборотами. Решение принято следующее — такие конструкции описываются и анализируются подобно типичным по специально созданным правилам. Поскольку единичное можно задать только перечислением, то достаточно гибкая система всегда позволит

пополнить этот список. Однако в целом грамматика должна быть системой правил, охватывающей бесконечно большое число структур входного языка.

### 7.7.6. Пример анализа и синтеза

Рассмотрим на примере английских предложений построение правил грамматики, описывающих структуры этих предложений. Все слова, входящие в предложение, анализируются без пропусков, независимо от того, имеют ли они переводной эквивалент, а также являются ли они частью устойчивого словосочетания.

Будем пользоваться следующими обозначениями:

S (sentence) — предложение, NP (noun phrase) — именная группа, VP (verb phrase) — глагольная группа, N (noun) — существительное или его заменитель, например личное местоимение, D (determiner) — артикль, притяжательное или указательное местоимение, Aux (auxiliary) — вспомогательный глагол, V (verb) — глагол, PP (prepositional phrase) — предложная группа, P (preposition) — предлог, AP (adjective phrase) — определительная группа, MV (main verb) — группа смыслового глагола, BE — обозначение всех форм глагола to be (am, is, are, was, were, been, being). В списке правил также должны быть и такие, которые определяют и пустую составляющую.

Такой грамматики уже достаточно для того, например, чтобы описать процесс распознавания предложений типа Peter is a pupil of the tenth form. Анализ такой структуры выглядит следующим образом:

1. S → NP VP
2. NP → D AP N, где D → @, AP → @, N > имя собственное и ИМЯ СОБСТВЕННОЕ → Peter
3. VP → Aux MV, где Aux → @
4. MV → BE NP PP, где BE → is
5. NP → D AP N, где D → a, AP → @
6. PP → P NP, где P → of
7. NP → D AP N, где D → the, AP → NUMB, NUMB → tenth, N → СУЩЕСТВ., СУЩЕСТВ.→ form.

Правилами такого рода можно описать лишь сравнительно небольшую часть английских предложений. Для создания более полной грамматики нужно осуществить предварительный отбор текстов или совокупности отдельных предложений по степени возрастания сложности их структур. Новую, более полную грамматику можно при этом разрабатывать «вглубь». Это позволяет расширять грамматику, включать в нее новые правила, охватывающие новые конструкции, возможные в данном языке.

Однородные члены анализируются с помощью рекурсивных правил грамматики. Если встречается несколько однородных членов, то для их анализа производится многократное применение грамматических правил. Так, предложение Swimming, fishing and walking are his favourite pastime анализируется с использованием следующих правил подстановок:

1. NP → NP, NP
2. NP → NP and NP
3. NP → NP NP
4. NP → Gerund
5. Gerund → swimming I fishing I walking

Кроме рекурсивных правил, при составлении грамматики учитываются еще и правила, оформляющие прерывистые составляющие, подобные тем, которые нужны, например для анализа следующих предложений:

Peter called Mary up.

They found the answer out.

В этих предложениях called... up и found... out — единые составляющие, разделенные на части словами Mary и the answer. Для подобных структур существуют правила, составленные по образцу:

S → NP VP
VP → Aux MV
MV → V Q P,

где Q есть элемент, образующий разрыв составляющей. В приведенных примерах Q может быть выражено различными типами NP.

Для того чтобы грамматика правильно анализировала предложение входного языка, необходимо соблюдение следующих требований:

1. Обеспечение возможно более полного охвата структур описываемого языка. Это возможно при условии, если тестовые предложения будут отбираться не случайно, а на основе тщательной выборки среди разнообразного текстового материала текстов в рассматриваемой предметно-тематической области. Полнота может быть обеспечена, если исходные тексты хорошо представляют генеральную совокупность. Кроме того, для построения правил грамматики может быть использован метод дедуктивных моделей, позволяющий предсказывать возможные структуры входного языка. В качестве исходного материала могут быть использованы учебники грамматики, специальные методические разработки, опыт создания анализирующих алгоритмов для других систем машинного перевода. На практике соблюдение требования полноты грамматики достаточно трудно осуществить, поэтому построенная грамматика и соответствующие алгоритмы должны быть достаточно гибкими в отношении создания и введения в грамматику новых правил для анализа неучтенных структур.
2. Принцип экономичности. Этот принцип предусматривает создание рационального количества правил. Если первое условие предусматривает полноту и всесторонний учет правил, то этот второй принцип направлен на то, чтобы избежать ненужной детализации, повторов и избыточности описания.
3. Непротиворечивость. Необходимо четкое разграничение подобных и разных структур. Правила грамматики не должны противоречить друг другу. Логичное следование одного из другого, непересечение одних правил с другими позволит избежать ошибок в работе анализатора.
4. Удобство и легкость записи правил. При минимальности условных символов грамматика должна нести максимум информации. А с учетом того, что грамматика будет расширяться, важно, чтобы она была понятна и доступна не только специалисту-лингвисту, принимавшему участие в

разработке правил анализа и синтеза текста, но и его возможному преемнику.

Данные требования, с 1 по 4, являются лингвистическими по содержанию. Следующее требование носит скорее комплексный — лингвистически-организационный характер:

5. Оптимизация алгоритмической обработки текста в соответствии с правилами грамматики. В эти правила должны быть заложены принципы, обеспечивающие наибольшую скорость анализа структуры предложения, вычисления атрибутов и генерации выходного текста.

Синтактико-семантические правила грамматики предназначены для построения дерева вывода соответствующей структуры анализируемого предложения.

### 7.7.7. Построение дерева вывода

Дерево вывода представляет собой схему анализа предложения сверху вниз, когда за исходную точку берется все предложение. Далее оно дробится на две и более составляющие. Каждая составляющая, в свою очередь, также разбивается на две и более части. Так продолжается до тех пор, пока в вершинах дерева не будут стоять только терминальные символы, т. е. слова входного языка.

Проиллюстрируем работу алгоритма построения дерева вывода для предложения Peter is a pupil of the tenth form.

Предположим, что входной язык более или менее полно описан и что список правил, анализирующих структуру его предложений, задан заранее.

Сначала к запрошенному предложению будет применено самое первое правило из списка, включающего все правила, а именно правило вида B → Y, где Y — составляющая S. Для указанного предложения будет использовано правило S → NP VP. Отметим, что если бы первоначально было взято другое правило для S, то алгоритм рано или поздно на каком-либо этапе отверг бы его как несостоятельное для данной структуры предложения. После этого стало бы проверяться другое правило, и так происходило бы до тех пор, пока не было бы найдено

единственно правильное решение. Иногда для того чтобы понять, какое правило должно быть применимо к данной структуре, необходимо практически построить все дерево вывода. Но чаще ошибочность правила определяется уже на более ранних этапах.

### 7.7.8. Задание правил вычисления атрибутов и правил выполнения генерации выходного текста

Создание и функционирование транслирующей системы автоматического перевода на базе инструментального комплекса ЛИНГВИК предусматривает также специальные правила вычисления атрибутов и генерации выходного текста. Эти два вида правил имеют сходную внешнюю форму представления и записываются в одно множество правил. Основное различие между ними связано с моментом их применения: правила вычисления атрибутов выполняются одновременно с построением дерева вывода (в момент присоединения к дереву вывода очередной новой вершины), а правила генерации выходного текста (правила семантического синтеза) выполняются после окончания построения дерева вывода, которое, фактически, управляет порядком их применения.

Правила вычисления атрибутов используются в основном для вычисления значений атрибутов и передачи их от терминальных вершин дерева к его корню. Необходимость использования правил вычисления атрибутов объясняется тем, что атрибуты исходных терминальных символов — слов английского языка — не всегда совпадают с атрибутами, требуемыми для правильного определения их переводных эквивалентов, что объясняется разницей грамматических систем английского и русского языков. Кроме того, правила вычисления атрибутов применяются тогда, когда в процессе синтактико-семантического анализа нужно выполнять ряд нетрадиционных действий, таких, например, как проверки контекстного окружения.

Оба указанных вида правил представляют собой последовательности операционных символов, каждый из которых предписывает транслирующей системе выполнить то или иное действие. Имеется возможность предусматривать разные по-

следовательности операционных символов в зависимости от результатов проверки некоторого заданного условия.

Развитие данной системы машинного перевода с целью достижения высококачественного машинного перевода, пригодного для постредактирования и экономически выгодного при этом, подразумевает развитие в следующих основных направлениях:

- пополнение автоматического словаря, полностью или в любом случае практически полностью покрывающего тематику данной предметной области;
- создание более полных описаний лексических значений слов;
- обеспечение гибкости всей системы в целом и словаря в частности для облегчения перехода на новые проблемно-тематические области;
- пополнение системы правилами лингвистической обработки, позволяющими получать на выходе понятные и легко редактируемые тексты;
- совершенствование включенных в систему описаний для повышения ее эффективности в части работы с этими описаниями;
- описание межфразовых связей в тексте, которое обеспечивало бы анализ текста как единого целого, а не как набора отдельных предложений, что дало бы возможность давать правильные переводы местоимений и обеспечивать передачу анафорических связей.

Совершенствование программного обеспечения могло бы также идти и в направлении улучшения связи с другими компьютерами.

Программная реализация системы САПФИР не была осуществлена вследствие того, что не было закончено составление лингвистического обеспечения системы. Как видно из вышеизложенного, основная трудоемкость работы заключается именно в создании лингвистического обеспечения. Тем не менее вышеизложенное представляется полезным, поскольку здесь с начала и до конца процесса осуществления машинного перевода четко перечислены все этапы, механизмы и лингвис-

тические составляющие этого сложного процесса. Как и в других подходах к реализации машинного перевода, основная часть работы заключается в детальном описании лингвистического содержания каждого из последовательных шагов алгоритма обработки входного текста для получения адекватного выходного текста.

## 7.8. Общая стратегия разработки систем машинного перевода на основе модели переводных соответствий

### 7.8.1. Общие принципы построения модели

Модель машинного перевода на основе переводных соответствий описана нами в (Марчук, 1983, 1985, 1997). Теоретический принцип, заложенный в эту модель, заключается в воспроизведении действий переводчика, работающего в данной языковой паре. Последовательно двигаясь от фразы к фразе, переводчик строит в уме некоторое приближенное представление о содержании текста, затем сопоставляет это представление с языковыми средствами, выбирая переводные эквиваленты и подыскивая переводные соответствия трех типов: эквивалентные, вариантные и трансформационные. Такая модель действий переводчика заимствована нами из традиционной теории перевода. Различаются также статика и динамика переводческого процесса. Статическая часть — словарь, правила грамматики, семантические закономерности. Динамическая часть — система нахождения переводных соответствий. Сами эти соответствия целесообразно находить и устанавливать на основе параллельных текстов, исходя из концепций современной корпусной лингвистики.

Как показал П. Н Хроменков (см. выше: Хроменков, 2000), все современные системы МП практического направления так или иначе действуют именно на основе концепции переводных соответствий, иногда в некоторых модификациях. В частности, такими модификациями являются так называемые траисферные системы, где переводные соответствия образуют особый этап, называемый трансфером (Новиков В. А., 2001). Совер-

шенствование качества перевода может зависеть от введения в алгоритмы анализа и синтеза более совершенных правил, например таких, которые разработаны для системы САПФИР, однако этот путь достаточно трудоемкий.

### 7.8.2. Блок-схема алгоритма МП на основе переводных соответствий

В настоящее время сложилось четкое представление о последовательности действий, направленных на реализацию модели переводных соответствий в конкретной языковой паре. В начале общего алгоритма производится анализ входного текста последовательным разбором входных словоформ. Этот анализ может быть назван автоматическим морфологическим анализом со всеми его характеристиками, которые описаны в соответствующем разделе настоящей книги. Отдельным этапом является выделение словаря неразложимых словосочетаний Далее производится анализ грамматической структуры предложений, причем содержание этого анализа определяется особенностями грамматики входного языка и необходимостью передачи этих особенностей в выходном языке. Перевод слов входного текста производится примерно посередине процесса обработки. Необходимость осуществления перевода до завершения процесса грамматического анализа объясняется тем, что ряд грамматических категорий выходного языка может быть определен только после перевода. Так, при переводе с английского языка на русский для правильного синтеза русских словоформ и предложения в целом необходимо знать род существительного. Категории рода в английском языке нет, поэтому грамматический анализ входного текста до перевода не может дать информацию о роде существительного. После перевода появляется возможность использовать информацию о роде для правильного оформления русской словоформы и для соответствующего оформления глагольных категорий как в морфологии, так и в синтаксисе. Последующие этапы анализа и синтеза дают возможность не только правильным образом перевести и дать соответствующую информацию каждой входной словоформе, но и постро-

ить предложение в соответствии с требованиями синтаксиса выходного языка.

Последовательность этапов анализа и синтеза текста в системе МП, построенной по модели переводных соответствий, имеет следующий вид:

1. Поиск словоформы в тексте и морфологический анализ.
2. Работа словаря оборотов.
3. Грамматический анализ до перевода.
4. Перевод однозначных слов.
5. Перевод многозначных слов.
6. Грамматический анализ после перевода.
7. Синтез выходных словоформ и текста.

*Рис. 20.* **Блок-схема общего алгоритма машинного перевода в модели переводных соответствий**

Стратегия реализации этой модели складывается из следующих действий:

1) подбор исходного корпуса текстов;
3) установление переводных соответствий на разных уровнях языковой структуры;
4) определение формальных категорий для морфологии, синтаксиса и семантики;
5) составление алгоритмов идентификации формальных категорий;
6) составление общего алгоритма системы машинного перевода.

В приведенной блок-схеме не указаны некоторые вспомогательные этапы, такие, как проверка правильности грамматической информации, этапы перестановки порядка слов и пр., которые зависят от типов сопоставляемых языков.

Данная схема была применена к разработке систем МП АМПАР, НЕРПА (немецко-русский автоматический перевод), СПРИНТ (Всесоюзный центр переводов). Кроме того, по этой схеме разрабатываются системы МП для таких языковых пар, как англо-персидский и русско-персидский МП (Али-Реза Ва-

липур, 1998; Мохаммад Реза, 1998; Эммарлу Рамезанали, 1998), а также МП с русского языка на китайский.

Данная общая схема, как показывает опыт, достаточно хорошо подходит к языкам различного строя. При этом следует учесть, что реализация ее должна проходить по общим законам корпусной лингвистики. Корпусная лингвистика, базирующаяся на больших массивах языковых текстов, позволяет извлекать объективные лингвистические данные и делать содержательные выводы на основе этих данных с высокой степенью достоверности, что гарантирует эффективность работы системы МП, построенной с учетом этих данных. При этом важен учет параллельных текстов, т. е. текстов оригинала и перевода.

Последнее особенно важно для МП. Первые работы с параллельными текстами были выполнены в конце 1980-х — начале 90-х годов в рамках создания различных систем статистического машинного перевода (Рахимбердиев, 2002; Gaussier et al., 2000). Была создана первая система машинного перевода, извлекающая знания о языке оригинала, языке перевода и правилах перевода исключительно из массивов примеров перевода. Эти работы вызвали критику со стороны «традиционной лингвистики», так как в них отвергались подходы к машинному переводу, считавшиеся общепринятыми в течение десятков лет. Несмотря на недостатки этих первых систем, статистический подход к МП привлек к себе внимание колоссальным сокращением трудоемкости построения таких систем по сравнению с системами традиционными. Неоспоримым достоинством таких систем является отказ от ручного составления машинных словарей и грамматик; если в логике системы обнаруживается ошибка, ее устранение в худшем случае означает необходимость повторного запуска процедуры извлечения параметров из корпуса примеров, а не ручное переписывание этих ресурсов.

Первые методики машинного перевода на основе примеров послужили толчком для дальнейших исследований в области извлечения лингвистической информации из параллельных текстов. С одной стороны, развивались сами системы статистического машинного перевода — за счет разработки более

сложных лингвистических моделей, отказа от идеи абсолютной независимости алгоритмов от обрабатываемых языков, привлечения морфологического анализа и т. п. Так, хотя в описанных системах статистического перевода не используются развитые формальные грамматики, понятие класса слов уже стало частью таких алгоритмов (см. также: Кривоносов, 2001). С другой стороны, произошла некоторая переоценка роли параллельных текстов. Алгоритмы обучения на параллельных корпусах стали применяться в инструментах, облегчающих труд переводчика, а также в системах автоматической проверки переводов, выполненных человеком (Марчук, 2002).

Таким образом, новая ветвь прикладной лингвистической науки, а именно корпусная лингвистика, которая имеет свою методику, цели и задачи обработки естественно-языковых текстов, вносит существенный вклад в создание эффективных современных систем машинного перевода.

## Заключение

Представление о составе и структуре науки компьютерной лингвистики еще не вполне установилось. Что именно из прикладной лингвистики входит в ее состав и структуру, пока не вполне ясно. Однако такое положение не является чем-то особенным для любой науки, и поэтому есть все основания не только говорить о компьютерной лингвистике, но и включать в нее в основном те разделы лингвистической науки, прикладной лингвистики, которые наиболее тесно связаны с компьютером, моделированием, программированием, искусственным интеллектом и другими проблемами современного информационного века.

Лингвистические основы компьютерной лингвистики можно строить по-разному. В вышедшем недавно в С.-Петербурге учебнике по прикладному языкознанию (Прикладное языкознание, 1996) материал организован, если можно так выразиться, по функциональному признаку: рассматриваются отдельные проблемы, главным образом, кардинальные, вокруг которых организуется прикладное языкознание и которые решаются с помощью специальных прикладных методов. Например, само-

стоятельные разделы этого учебника названы «Орфография», «Моделирование языка», «Автоматизированные обучающие системы», «Психолингвистика», «Учебная лексикография», «Статистическая обработка экспериментальных данных» и пр.; таким образом, в качестве самостоятельных компонентов науки выступают методы, предметные области, языковые аспекты, результаты и т. п. В отличие от этого в нашем изложении мы последовательно придерживаемся принципа движения от наименьшей лингвистической единицы (буквы, слова) ко все более сложным единствам языка в аспектах прикладного языкознания и во взаимодействии с информатикой. Нам представляется, что таким образом более четко вырисовывается именно лингвистическое содержание информатических основ в тех проблемных областях, в которых сейчас работает прикладное языкознание и компьютерная лингвистика.

Думается, что наиболее существенной чертой лингвистических основ информатики и прикладной лингвистики в целом является возможность строить и проверять на фактическом материале работу воспроизводящих инженерно-лингвистических моделей (или как их можно назвать по-другому, моделей с обратной связью). Любая гипотеза, даже весьма абстрактная, может быть проверена с помощью таких моделей и по результатам такой проверки либо принята, либо скорректирована. Вряд ли какая-либо другая лингвистическая наука располагает такими возможностями практической проверки и совершенствования теории.

Продолжается ускоренное развитие технических средств и компьютерных возможностей, которые, безусловно, будут и далее значительно влиять на развитие лингвистических средств и теорий, связанных с информатикой. Социальная потребность в решении многих комплексных задач, таких, например, как перевод, продолжает также возрастать и по крайней мере не имеет перспектив быть удовлетворенной без решения соответствующих лингвистических проблем. Все более широкий лингвистический спектр будут приобретать задачи искусственного интеллекта.

Несмотря на то, что язык как некоторая предметная среда, в которой происходит коммуникация, гораздо более консерва-

тивен, чем среда информатическая, и несмотря на постулируемое иногда естественно-искусственное многоязычие, происходит движение к сближению информатических проблем и проблем языковедения имеет место, но оно в значительно большей степени исходит от информатики, чем от лингвистики. Тем не менее практика развития человеческого языкового общения и развитие теории позволяют надеяться на то, что уже в ближайшем будущем нас ждут новые интересные открытия в сопредельной области, краткому описанию которой и посвящена настоящая книга.

Мы отдаем себе отчет в том, что далеко не все проблемы компьютерной лингвистики рассмотрены достаточно подробно в настоящей монографии. Как теория, так и особенно практика этой науки развиваются быстрыми темпами. Быстро растут и появляются новые возможности современных информационных технологий. Однако можно считать, что к настоящему времени определились некоторые основы этой науки — компьютерной лингвистики, — которые мы и попытались изложить в этой книге. Как и всякие основы, они тоже не застрахованы от пересмотра, расширения, дополнений и изменений. Но, по крайней мере на сегодняшний день, то, что здесь рассматривается как основы и главное содержание компьютерной лингвистики, прошло проверку временем и практикой создания и эксплуатации действующих систем автоматической обработки естественно-языковых текстов. Пусть данная книга послужит хоть небольшим вкладом в развитие и разработку соответствующей интересной и перспективной проблематики.

# Литература


Авербух К. Я. Общая теория термина. Иваново, Ивановский госуниверситет, 2004. 251 с.

Агеев В. Н. Электронная книга: метафора или новая реальность? // Человеческий фактор в правоохранительных системах. Материалы Международной конференции. Орел, 1996. С. 284–296.

Агеев В. Н., Узилевский Г. Я. Человеко-компьютерное взаимодействие: концепции, процессы, модели. М., Мир книги, 1955. 352 с.

Али-Реза Валипур. Анализ и синтез глагольных форм и конструкций при машинном переводе с русского языка на персидский. Канд. дисс. М.: МГУ, 1998. 121 с.

Андреева Е. С. Диалектика текста. Опыт логико-лингвистического синтеза. М., УРСС, 2001. 96 с.

Бартков И. И. Корреляционный анализ в дериватологии. // Дериватология и дериватография литературной нормы и научного стиля. Владивосток: АН СССР, ДВНЦ, 1984. С. 3–27.

Бартков В. И. Количественная морфемография (дериватография) английского, немецкого, французского и русского языков (научный стиль и литературная норма). // Основосложение и полуаффиксация в научном стиле и литературной норме. Владивосток, АН СССР, ДВНЦ, 1982. С. 27–55.

Белоногов Г. Г., Хорошилов Ал-др, Гуськова Л. Ю., Хорошилов Ал-сей, Козачук М. В., Рыжова Е. Ю. Каким быть машинному переводу в XXI веке. В кн: Перевод: традиции и современные техиологии. М.: ВЦП, 2002. С. 56–69.

Белоногов Г. Г., Новоселов А. П. Автоматизация процессов накопления, поиска и обобщения информации. М.: Наука, 1979. 255 с.

Белоногов Г. Г. Определение грамматических признаков «новых» слов с помощью словаря. // Инженерная лингвистика.

Л., 1971, Уч. записки ЛГПИ им. А. И. Герцена. Т. 458. Ч. II. С. 225–229.

Бельская И. К. Язык человека и машина. М.: Изд-во МГУ, 1969. Т. 1. 408 с. Т. 2. 250 с.

Борисова Е. Г. Принципы описания коллокаций (на материале русского языка). Автореф. докторской дисс. М., МГЛУ, 1996. 24 с.

Борисова Л. И. Ложные друзья переводчика. Общенаучная лексика. Английский язык. М.: НВИ Тезаурус, 2002. 211 с.

Брябрин В. М. Программное обеспечение персональных ЭВМ. М.: Наука, 1988. 272 с.

Вайнцвайг М. Н., Полякова М. П. Механизм мышления и моделирование его работы в реальном времени. // Интеллектуальные процессы и их моделирование. М.: АН СССР, 1987. С. 208–229.

Варга Д. Проблемы осуществления морфологического анализа при машинном переводе. Научно-техническая информация (НТИ) М.: ВИНИТИ, 1964. № 4. С. 47–50.

Василевский А. Л., Марчук Ю. Н. Вычислительная лингвистика. Учебное пособие для студентов отделения прикладной лингвистики. М.: МГПИИЯ им. М. Тореза, 1970. 265 с.

Веденов А. А. Моделирование элементов мышления. М.: Наука, 1988. 159 с.

Вертель Е. В. Проблемы машинной лексикографии. Дис. канд. филол. наук. М.: МГУ, 1984. 125 с.

Власов В. К., Королев Л. Н., Сотников А. Н. Элементы информатики. М.: Наука, 1988. 318 с.

Галяшина Е. И. Основы судебного речеведения. М.: СТЭНСИ, 2003.

Головин Б. Н. Введение в языкознание. М.: Высшая школа, 1966. 332 с.

Гринев С. В. Введение в терминографию. М.: МПУ, 1995. 158 с.

Гуревич П. С. Культурология. М.: Знание, 1996. 287 с.

Денисов П. Н. Принципы моделирования языка. М.: МГУ, 1965. 151 с.

Диалог-97. Труды международного семинара по компьютерной лингвистике и ее приложениям. Ясная Поляна, 10–15 июня 1997. 315 с.

Диалог-96. Труды международного семинара по компьютерной лингвистике и ее приложениям. Пущино, 4–9 мая 1996. 305 с.

Диалог-95. Труды международного семинара по компьютерной лингвистике и ее приложениям. Казань, 31 мая — 4 июня 1995. 362 с.

Занегина Н. Н. Многозначные слова в многоязычных терминологических словарях. Вестник МГЛУ. Серия 1. Филология. № 7/2001. Минск: МГЛУ, 2001. С. 125–140.

Заслонко А. Н., Хахалин Г. К. Лингвистический транслятор для системы машинного перевода // Международный семинар по машинному переводу. Тбилиси, 1989. С. 111–113.

Звегинцев В. А. Предложение и его отношение к языку и речи. М.: Изд-во МГУ, 1976. 306 с.

Зеленков Ю. Г. Морфологический анализ в системах автоматической обработки научно-технической информации. Канд. дис. М.: ВИНИТИ, 1988. 145 с.

Зелко В. М. Проблемы разработки лингвистического обеспечения системы китайско-русского информационного машинного перевода. Канд. дис. М.: Ин-т языкознания АН СССР, 1991. 165 с.

Зубов А. В., Зубова И. И. Информационные технологии в лингвистике. М.: ACADEMIA, 2004. 205 с.

Зубов А. В. (отв. ред.) Компьютерная лингвистика и обучение языкам. Сборник статей. Минск: МГЛУ, 2000. 221 с.

Зубова И. И. Информационные технологии в лингвистике. Минск: МГЛУ, 2001. 211 с.

Ершов А. П. Машинный фонд русского языка: внешняя постановка. // Машинный фонд русского языка: идеи и суждения. М.: Наука, 1986. С. 7–12.

Естественный язык, искусственные языки и информационные процессы в современном обществе. Отв. ред. Котов Р. Г. М.: Наука, 1988. 135 с.

Ефимов Н. Н., Фролов В. С. Основы информатики. Введение в искусственный интеллект. М.: МГУ, 1991. 115 с.

Кан Д. Взломщики кодов. М.: Центрполиграф, 2000. 473 с.

Караулов Ю. Н. Методология лингвистического исследования и машинный фонд русского языка. // Машинный фонд русского языка: идеи и суждения. М.: Наука, 1986. С. 13–25.

Карпов В. А. Язык как система. Минск: Вышэйшая школа, 1992. 302 с.

Князев С. В. Еще раз к вопросу о соотношении фонетики и фонологии. Вестник Московского Университета. Серия 9. Филология. 6/2001. С. 101–111.

Кобрин Р. Ю. Лингвистическое описание терминологии как база концептуального моделирования в информационных системах. Автореф. докторской дисс. Ленинград, 1989. 42 с.

Козловский С. В. Лингвистические процессоры для персональных экспертных систем // Проблемы автоматического и экспертно-фонетического анализа текстов. Минск, 1986. С. 67–170.

Козьмина Е. Л. Обратные словари. Принципы их создания и использования. Канд. дис. М.: МГУ, 1988. 152 с.

Колегов А. В, Международный язык-посредник эльюнди. Тирасполь: Папирус, 2003. 496 с.

Колодяжная Л. И. Автоматизированная лексикографическая система УНИЛЕКС. М.: МГУ, 1987. 115 с.

Комиссаров В. Н. Слово о переводе. М.: ИМО, 1973. 237 с.

Корнилов О. А. Языковые картины мира как производные национальных менталитетов. М.: ЧеРо, 2003. 347 с.

Королев Э. И. Промышленные системы машинного перевода. М.: Всесоюзный Центр переводов, 1991. 104 с.

Коростелев Л. Ю. Стохастический подход к графематике слов (проблема диагностики и коррекции искажений в системе

машинного перевода). Канд. дис. Военный Краснознаменный институт, 1985. 135 с.

Коростелев Л. Ю., Марчук Ю. Н. Анализ слов, отсутствующих в словаре, в системах автоматизированной обработки текстов. // Вопросы кибернетики: прикладные аспекты лингвистической теории. М.: АН СССР, 1987. Вып. 115. С. 103–116.

Котов Р. Г., Новиков А. И., Скокан Ю. П. Прикладная лингвистика и информационная технология. М.: Наука, 1987. 163 с.

Котов Р. Г., Домбровская И. В., Скокан Ю. П. Один из вопросов коммуникации «человек-машина». // Проблемы бионики. Харьков: Вища школа, 1988. Вып. 40. С. 18–23.

Кривоносов А. Т. Система классов слов как отражение структуры языкового создания. Москва — Нью-Йорк: Че-Ро, 2001. 846 с.

Кривоносов А. Т. Язык, логика, мышление. Умозаключение в естественном языке. Москва — Нью-Йорк: Валанг, 1996. 682 с.

Лебедев А. Н. Циклические коды памяти // Когнитивная психология. М.: Наука, 1986. С. 106–115.

Лесников С. В. Русский гипертекстовый тезаурус (Гизаурус). В кн: Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных». СПб: Изд-во СПб Университета, 2002. С. 95–97.

Лесохин М. М., Лукьяненков К. Ф., Пиотровский Р. Г. Введение в математическую лингвистику. Минск: Наука и техника, 1982. 263 с.

Лингвистическая прагматика и общение с ЭВМ. Отв. ред. Ю. Н. Марчук. М.: Наука, 1989. 141 с.

Лингвистические вопросы алгоритмической обработки сообщений. Ред. Котов Р. Г., Курбаков К. И. М.: Наука, 1983. 246 с.

Лингвистический энциклопедический словарь. М.: Советская Энциклопедия, 1990. 683 с.

Линс Ульрих. Опасный язык. Книга о преследованиях эсперанто. М.: Права человека, 1999. 574 с.

Макаров Н. П. Полный русско-французский словарь. СПб, 1908, изд. Макарова Н. П., 1127 с.

Маковский М. М. Лингвистическая комбинаторика. М.: Наука, 1988. 232 с.

Максименко О. И. Формальные методы в современной прикладной лингвистике. М.: МГОУ, 2002. 256 с.

Мартыненко Г. Я. Методы статистического моделирования в языкознании. // Прикладное языкознание. СПб: Изд-во СПб университета, 1996. С. 201–223.

Марчук М. В. Динамика лексических значений многозначных слов (лексика основного терминологического слоя). Дис. в форме научного доклада на соиск. уч. степени доктора филол. наук. М.: МПУ, 1996. 59 с.

Марчук М. В. К развитию лексических значений многозначных слов. Канд. дисс. Л.: ЛГУ, 1988. 125 с.

Марчук Ю. Н. Русская терминология в современной языковой ситуации. В сб. Язык и речь: проблемы и решения. М.: МГУ им. М. В. Ломоносова, 2004. С. 347–356.

Марчук Ю. Н. (2004). Универсальный язык — путь к взаимопониманию и миру. Мир нравственности. М.: Межрелигиозная межнациональная организация за мир во всем мире. № 10. 2004. С. 16–17.

Марчук Ю. Н. Машинный перевод с русского языка на восточные языки (некоторые аспекты). В кн: Перевод: традиции и современные технологии. М.: ВЦП, 2002. С. 70–75.

Марчук Ю. Н. Корпус текстов и сверхбольшие базы лингвистических данных. Доклады научной конф. «Корпусная лингвистика и лингвистические базы даных». СПб: Изд-во СПб Университета, 2002. С. 102–107.

Марчук Ю. Н. Компьютерная лингвистика и ее приложения. Научно-аналитические обзор. М.: ИНИОН РАН, 1998. 125 с.

Марчук Ю. Н. Модель «текст–текст» и переводные соответствия в теории машинного перевода. // Проблемы компьютерной лингвистики. Минск: МГЛУ, 1997. С. 21–29.

Марчук Ю. Н. Терминологическая работа Всероссийского центра переводов // Терминоведение. М.: Московский Лицей, 1996. Вып. 1–3. С. 14–18.

Марчук Ю. Н. Теория и практика машинного перевода. // Русский филологический вестник. М.: Московский Лицей, 1996. Т. 81. С. 123–135.

Марчук Ю. Н. Основы терминографии. М.: ЦИИ МГУ, 1992. 75 с.

Марчук Ю. Н. Математические методы в языкознании. Актуальные проблемы прикладного языкознания. М.: ИНИОН АН СССР, 1990. 47 с.

Марчук Ю. Н. Методы моделирования перевода. М.: Наука, 1985. 203 с.

Марчук Ю. Н. Проблемы машинного перевода. М.: Наука, 1983. 232 с.

Марчук Ю. Н. Вычислительная лексикография. М.: ВЦП, 1976. 183 с.

Марчук Ю. Н. (отв. ред). Контекстологический словарь для машинного перевода многозначных слов с английского языка на русский. М.: ВЦП, 1976. Ч. 1. 264 с. Ч. 2. 256 с.

Марчук Ю. Н. Опыт машинной реализации дистрибутивной методики определения лексических значений. // Статистика речи и автоматический анализ текста. 1972. Л.: Наука, 1973. С. 181–230.

Марчук Ю. Н., Моторин Ю. А. Основные принципы автоматизации перевода с английского языка на русский. Вопросы радиоэлектроники. МРП СССР, серия ЭВТ, 1970. Вып. 7. С. 11–19.

Маслов Ю. С. Введение в языкознание. М.: Высшая школа, 1987. 271 с.

Мельчук И. А., Равич Р. Д. Автоматический перевод 1949–1963. М.: ВИНИТИ, 1967. 517 с.

Мельников Г. П. Системология и языковые аспекты кибернетики. М.: Советское Радио, 1978. 367 с.

Меркулова С. В. Толковый словарь страховых терминов. Dictionary of Insurance Terms. М.: Изд-во МАИ, 2000. 135 с.

Милославский И. Г. Краткая практическая грамматика русского языка. М.: Русский язык, 1987. 283 с.

Михайлов А. И., Черный А. И., Гиляревский Р. С. Основы информатики. М.: Наука, 1968. 231 с.

Михайлов В. Г. Параметрические модели речевой связи (системное описание физико-технических параметров речи) // Некоторые аспекты формирования базы лингвистических знаний для автоматизированных систем. М.: МГЛУ. Деп. в ИНИОН РАН, № 50875, 2.11.95. С. 15–27.

Михайлов В. Г., Златоустова Л. В. Измерение параметров речи. М.: Радио и связь, 1987. 167 с.

Моторин Ю. А., Марчук Ю. Н. Реализация автоматического перевода на современных серийных ЭВМ общего назначения. — Вопросы радиоэлектроники. — МРП СССР, серия ЭВТ, 1970. Вып. 7. С. 20–39.

Мохаммади Мохаммад Реза. Система русско-персидского машинного перевода на основе переводных соответствий. Канд. дисс. М.: МГУ, 1998. 132 с.

Научно-технический перевод. Сб. статей под ред. Ю. Н. Марчука. М.: Наука, 1987. 125 с.

Наседкин А. Д. Язык как средство формирования мысли.// Язык и мышление. М.: Наука, 1967. С. 65–73.

Нелюбин Л. Л. Инженерно-лингвистическое моделирование и компьютерная лингвистика. // Проблемы компьютерной лингвистики. Минск: МГЛУ, 1997. С. 18–21.

Нелюбин Л. Л. Компьютерная лингвистика и машинный перевод. М.: Наука, 1983. 241 с.

Новиков А. И. Доминантность и транспозиция в процессе осмысления текста. В кн: Проблемы прикладной лингвистики 2001. М.: РАН, 2002. С. 155–180.

Новиков В. А. Трансфер в современных системах машинного перевода. М., канд. дисс. МПУ, 2001. 151 с.

Павловска Е. Ю. Исследование динамических характеристик баз данных для оценки тенденций развития новых научных направлений. Дисс. канд. техн. наук. М.: ВИНИТИ, 1988. 154 с.

Пальм Р. О морфологическом анализе русской фразы // Сообщения по машинному переводу. Таллин, 1962. Вып. 1. С. 59–83.

Петров В. В., Переверзев В. Н. Обработка языка и логика предикатов. Новосибирск: Изд-во Новосибирского Университета, 1993. 157 с.

Петров В. В. Язык и логическая теория // Новое в зарубежной лингвистике. М.: Прогресс, 1986. С. 5–23.

Пиотровский Р. Г. Лингвистический автомат и его речемыслительное обоснование. Минск: МГЛУ, 1999. 196 с.

Пиотровский Р. Г. Инженерная лингвистика и теория языка. Л.: Наука, 1979. 111 с.

Пиотровский Р. Г., Бектаев К. Б., Пиотровская А. А. Математическая лингвистика. М.: Высшая школа, 1977. 383 с.

Поляков В. Н. Использование технологий, ориентированных на лексическое значение, в задачах поиска и классификации. В кн: Проблемы прикладной лингвистики 2. М.: ИЯ РАН, 2004. С. 101–117.

Поляков С. Э. Мифы и реальность современной психологии. М., УРСС, 2004. 484 с.

Поспелов Д. А. О «человеческих» рассуждениях в интеллектуальных системах. Логика рассуждений и ее моделирование. // Вопросы кибернетики. М.: АН СССР, 1983. С. 5–37.

Потапова Р. К. Новые информационные технологии и лингвистика. М.: МГЛУ, 2002. 575 с.

Потапова Р. К. Речь: коммуникация, информация, кибернетика. М.: Радио и связь, 1997. 527 с.

Потапова Р. К. Технологии обработки естественного языка в науке и промышленности. Обзор. М.: ИНИОН РАН, 1992. 63 с.

Потапова Р. К. Речевое управление роботом. М.: Радио и связь, 1989. 247 с.

Прикладное языкознание. Учебник / Л. В. Бондарко, Л. А. Вербицкая, Г. Я. Мартыненко и др. Отв. ред. А. С. Герд. СПб: Изд-во СПб Университета, 1996. 528 с.

Проблемы прикладной лингвистики 2001. Отв. ред. Новиков А. И. М.: РАН, Ин-т языкознания, 2002. 359 с.

Прохорова В. Н. Русская терминология (лексико-семантическое образование). М.: МГУ, 1996. 125 с.

Рахимбердиев Б. Н. Эволюция семантики экономической терминологии русского языка в XX веке. Канд. дисс. М.: МГЛУ, 2003. 151 с.

Рогожникова Р. П. Машинный фонд русского языка и словарное дело. // Машинный фонд русского языка: идеи и суждения. М.: Наука, 1986. С. 58–67.

Рождественский Ю. В. Философия языка. Культуроведение и дидактика. М.: Грантъ, 2003. 239 с.

Рождественский Ю. В. Введение в общую филологию. М.: Высшая школа, 1979. 223 с.

Рождественский Ю. В. Типология слова. М.: Высшая школа, 1969. 321 с.

Рождественский Ю. В., Волков А. А., Марчук Ю. Н. Введение в прикладную филологию. М.: МГУ, 1988. 116 с.

Русская грамматика. Гл. ред. Шведова Н. Ю. М.: Наука, 1980. Т. 1. 783 с.

Рябцева Н. К. Язык и естественный интеллект: конгруэнтность, асимметрия, дополнительность. В кн. Проблемы прикладной лингвистики 2. М.: РАН, 2004. С. 71–100.

Рябцева Н. К. Лингвистическое моделирование естественного интеллекта и представление знаний. В кн: Проблемы прикладной лингвистики 2001. М.: РАН, 2002. С. 228–252.

Рябцева Н. К. Информационные процессы и машинный перевод. М.: Наука, 1986. 167 с.

Светова С. Ю., Косматова Е. В. Системы автоматизированного перевода PROMPT. Системы Translation Memory Trados.

Интеграция Trados и PROMPT. В кн: Перевод: традиции и современные технологии. М.: ВЦП, 2002. С. 42–55.

Седов А. Е. Тексты и формы — в живом и о живом: корни современной биологии в памятниках различных культур. // Русский исторический вестник. М.: Московский Лицей, 1998. С. 121–136.

Семенов А. Л. Контекстологический словарь основных терминов маркетинга. М.: ВЦП, 1994. 122 с.

Сифоров В. И., Чаповский А. З. Развитие работ в области научно-технической терминологии на современном этапе НТР //Теория и практика научно-технической лексикографии. М., 1988. С. 11–15.

Скокан Ю. П. Формирование знаний в компьютере на основании содержания смыслообразующих участков текста. В кн.: Проблемы прикладной лингвистики 2001. М.: РАН, 2002. С. 253–260.

Соболева Т. А., Суперанская А. В. Товарные знаки. М.: Наука, 1986. 172 с.

Советский энциклопедический словарь. М.: Сов. Энциклопедия, 1980. 1600 с.

Сухотин Б. В. Выделение морфем в текстах без пробелов между словами. М.: Наука, 1984. 96 с.

Сухотин Б. В. Оптимизационные методы исследования языка. М.: Наука, 1976. 157 с.

Танасиенко А. В. О некоторых методологических принципах построения систем машинного перевода нетрадиционного типа, базирующихся на обобщенно-модельной концепции лингвистического знака // Международный семинар по машинному переводу. Тбилиси, 1989, ВЦП. С. 299–302.

Татаринов В. А. Теория терминоведения: В 3 т. Т. 1. Теория термина: история и современное состояние. М.: Московский Лицей, 1996. 311 с.

Татаринов В. А. История отечественного терминоведения. М.: Московский Лицей. Т. 1. 1994. 407 с.; Т. 2. 1995. 333 с.

Теория и практика английской научной речи. М.: МГУ, 1987. 240 с.

Тулдава Ю. Проблемы и методы квантитативно-системного исследования лексики. Таллин: Валгус, 1987. 204 с.

Фитиалов С. Н. О построении формальной морфологии в связи с машинным переводом. Тезисы конф. по обработке информации, машинному переводу и автоматическому чтению текста. ВИНИТИ АН СССР. М., 1961. 21 с.

Французско-англо-русско-кхмерский географический словарь. Под ред. Марчука Ю. Н. и Чой Ауна. М.: ЦИИ МГУ, 1991. 293 с.

Французско-англо-русско-кхмерский лингвистический словарь. Под ред. Марчука Ю. Н и Со Муй Кхеанга. М.: ЦИИ МГУ, 1991. 203 с.

Харпер К. Предварительные исследования русского языка. // Машинный перевод. М.: Изд-во иностр. лит-ры, 1957. С. 101–124.

Хомутов А. В. Разработка и внедрение технологического комплекса обработки объектографической информации в центральном звене межотраслевой автоматизированной информационной системы. Канд. дис., М.: ВИМИ, 1989. 124 с.

Хроменков П. Н. Анализ и оценка эффективности современных систем машинного перевода. Канд. дисс. М.: МПУ, 2000. 126 с.

Чикобава А. С. К вопросу о взаимоотношении мышления и речи в связи с ролью коммуникативной функции. // Язык и мышление. М.: Наука, 1967. С. 16–30.

Шабанов-Кушнаренко Ю. П. Теория интеллекта. Математические средства. Харьков: Вища школа, 1984. 141 с.

Шаховской В. И. Эмотивность и лексикография. // Филологические науки, 1986. № 6. С. 42–47.

Шевчук В. Н. Динамика развития отраслевой терминологии как лексикографическая проблема // Теория и практика научно-технической лексикографии. М.: Русский язык, 1988. С. 57–61.

Шелов С. Д. Опыт построения терминологической теории: значение и определение терминов. Дис. докт. филол. наук. М.: МГУ, 1995. 201 с.

Шемакин Ю. И. Введение в информатику. М.: Финансы и статистика, 1985. 189 с.

Шемякина А. В. Компьютерный переводной словарь Мультилекс 3.5. В кн: Перевод: традиции и современные технологии. М.: ВЦП, 2002. С. 98–109.

Широков О. С. Языковедение. Введение в науку о языках. М.: Добросвет, 2003. 734 с.

Штиндлова Й. Обратные словари.//Автоматизация в лингвистике. Л.: Наука, 1966. С. 84–91.

Шуклин Д. Е. Морфологический и синтаксический разбор текстов как конечный автомат, реализованный семантической нейронной сетью, имеющей структуру синхронизированного линейного дерева. В кн: Новые информационные технологии. М.: МГИЭИМ, 2002. С. 74–85.

ЭВРИКА. Технологическое возрождение Европы: французские предложения от июня 1985 года. М.: ВЦП, 1987. 52 ср.

Эммарлу Рамезанали. Анализ и синтез именных групп при машинном переводе с русского языка на персидский. Канд. дисс. М.: МГУ, 1998. 127 с.

Юзвишин И. И. Информациология — научная основа построения информационной модели мира // Докл. на четвертом международном форуме информатизации. М.: 1995. 121 с.

Яглом А. М., Яглом И. М. Вероятность и информация. М.: Гос. изд. техн. теор. лит-ры, 1957. 159 с.

Altmann G. Science and Linguistics.//Contributions to Quantitative Linguistics. Dordrecht etc. Kluwer Acad. Publ.,1993, pp.3-11.

Arranz M. V., Radford I., Ananiadou S., Tsujii J. Towards a Sublanguage-Based Semantic Clustering Algorithm // Recent Advances in Natural Language Processing. Ed. by R. Mitkov & N. Nikolov. John Benjamins Publ. Co., Amsterdam/ Philadelphia, 1997, pp. 125-136.

Aslanides S., Danlos L. Génération d'un texte à partir d' un graphe événementiel dans le formalisme TAG. // Colloque «Informatique et Langue Naturelle», I. L. N '93, Nantes, 1993, pp. 27-58.

Bertaux, P. Les deux langages: analogique et digital. Paris, Didier Erudition, 1984, 89 pp.

Better Translation for Better Communication. G. Van Slype, J. F. Guinet,. Seitz, S. Benejam. Pergamon Press, Oxford etc, 1983, 141 pp.

Bilingual Lexicography in Poland: Theory and Practice. Ed. by Jan Wawrzynczyk. — Warszawa, Univ. Warszawski, — 1995, — 137 pp.

Boitet, Ch. Twelve Problems for Machine Translation. International Conference on Current Issues in Computational Linguistics. Universiti Sains Malaysia, Penang, Malaysia, 1991, Proceedings, pp. 45-57.

Carson J. Unification and Transduction in Computational Phonology.// Proceedings of COLING, Budapest, 1988, vol.1, pp. 106-111.

Constraints, language and computation /Ed. by Rupp C. J. et al. — L. etc.: Acad. Press, 1994. — 391 pp.

Denenberg R. Open Systems Interconnection. // Encyclopedia of Library and Information Science, vol.44, suppl.9, Marcel Dekker< Inc., — New-York-Basel, 1989, pp.210-232.

Descl`es J-P., Jouis Chr. L`exploration contextuelle: une me`thode linguistique et informatique pour l`analyse automatique de textes. // Colloque «Informatique & Langue Naturelle», Actes, Nantes, 1993, pp. 349-364.

Descriptive Tools for Electronic Processing of Dictionary data //Studies in Computational Lexicography. The DANLEX group: Hiorth E., Madsen B. N., etc. — Tuebingen, Niemeyer,1987. — 285 p.

Dubey Y. P. Decision Support Systems: Development and Trends. //Encyclopedia of Computer Science and Technology. Ed. A. Kent, J. G. Williams. Marcel Dekker, Inc., N. Y., Basel, 1988, pp.114-173.

Dugas A., Labelle D. Le groupe nominal № 1 de № 2 et autres suites N de N. // Informatique & Langue Naturelle, I. L. N. 93, Actes. Dec. 2-3, 1993, Nantes. Universite de Nantes,1993, pp. 445-456.

Early Years in Machine Translation. Ed. by W. John Hutchins. John Benjamins Publ. Co., Amsterdam/Philadelphia, 2000, 401 pp.

Garnier G. Linguistique et traduction: element de systematique verbale comparee du francais et de l'anglais. Caen; Paradigme, 1985, 505 pp.

Gaussier E., Hull D., Ait-Mokhar S. Term Alignment in Use: Machine-Aided Human Translation. In: Parallel Text Processing, ed. by J. Veronis, Dordrecht, Netherlands, 2000, pp. 121-153.

Herdan G. Language as Choice and Chance. P. Noordhoff, Groningen, 1956, 350 pp.

HeSS K., Brustkern J., Lenders W. Maschinenlesbare Deutsche Woerterbuecher. Max Niemeyer Verlag, Tuebingen. 1983-228 s.

Hill J. C. Language Acquisition // Encyclopedia of Artificial Intelligence. Vol. 1. New-York, E. A. 1987, pp.471-629.

Hutchins W. J. Machine Translation: Past, Present, Future. — Chichester, Ellis Horwood N. Y. etc. Wiley, 1986, 382 pp.

Kahn D. The Codebreakers. (The Story of Secret Writing). New-York, The Macmillan Co., 1968, 1164 pp.

Kokkinakis G. Electronic Dictionaries Integrating Multimedia and Speech Language Technologies. In: SPECOM'2001 Proceedings. Moscow, 29-31 Oct., 2001, pp.6-8.

Kuehlwein W. The need for Integration of Applied and Theoretical Linguistics.// The Relation of Theoretical and Applied Linguistics. Ed. by O. Miseska Tomic and R. W. Shuy. Plenum Press, N. Y. and London, 1987, 193 pp. p.51-73.

Marchuk Y. N. The Burdens and Blessings of Blazing the Trail. In: Journal of Quantitative Linguistics. Trier, Swets & Zeitlinger, Vol. 10, No. 2, Aug. 2003. pp 81-87.

Marchuk Yu. N. Machine-Aided Translation. A Syrvey of Current Systems // Computational Linguistics. Ein internationales Handbuch zur computergestutzten Sprachforschung und ihrer Anwendungen. Ed. by I. Batori et al. Walter de Gruyter, Berlin-New-York, 1989, pp. 682-688.

Marchuk Yu. N. Machine Translation in the USSR // Encyclopedia of Library and Information Science. Vol.44, suppl.9. M. Dekker Inc. N. Y. — Basel, 1989, pp. 183-194.

Marchuk Y. N. Machine Translation in the USSR. // Computers and the Humanities. Paradigm Press, Osprey, Florida, USA, 18 (1984), pp. 39-46.

Marchuk Y. N., Tihomirov B. D., Scerbinin V. I. Ein System zur maschinellen Uebersetzung aus dem Englischen ins Russische // Automatische Sprachuebersetzung. Wissenschaftliche Buchgesellschaft, Darmstadt, 1982, SS. 319-336.

Marchuk Y. N. The Contextological Dictionary: Use in Programmed Language Teaching. In: Computers and the Humanities 13 (1979), pp. 277-281.

Martynenko G. Semiotics of Statistics. In: Journal of Quantitative Linguistics, Trier, Swets & Zeitlinger, Vol.10, No. 2, Aug. 2003, pp. 105-116.

Mehrak Rahiniashtiani. Teaching English as a Foreign Language. M., Народный Учитель, 2002, 201 стр.

Miram G/E/ Translation Algorithms. Kyiv, «Twin Inter», 1998, 175 pp.

Oettinger A. Automatic Language Translation. Harvard Univ. Press, Cambridge, Mass., 1960, 380 pp.

Penguin Dictionary of Human Geography. Brain Goodall, Penguin Books, Harmondsworth, England, 1987. 509 pp.

Prospects for a New Structuralism. Univ. of Ottawa. Amsterdam-Philadelphia. — Benjamins,1992, 276 pp. Introduction.

Rentzepopoulos P. A., Tsopanoglou A. E., Kokkinakis G. K. A Statistical Approach for Phoneme-to-Grapheme Conversion.// Contributions to Quantitative Linguistics, Ed. by R. Koehler & B. Rieger. Kluwer Acad. Pubi., Dordrecht etc, 1993, pp.319-330.

Silnitsky G. Correlation of Phonetic and Morphological Systems of Indo-European Languages/ Journal of Quantitative Linguistics, Trier, Swets & Zeitlinger, Vol.10, No.2, Aug. 2003, pp. 129-142.

Tsujii J. Machine Translation: Productivity and Conventionality of Language. // Current Issues in Linguistic Theory — 136. Recent Advances in Natural Language Processing. Ed. by R. Mitkov et al. Lohn Benjamins Publ. Co., Amsterdam/Philadelphia, 1997, pp. 377-392.

Tsutsumi J., Nitta T., Ono K., Nobesawa Sh., Nakanishi M. Multi-Lingual Machine Translation Based on Statistical Information. QUALICO-94, 2nd Intern. Conf. on Quantitative Linguistics. Moscow Lomonossov State Univ., 1994, Proceedings, pp.147-152.

Wiener N. Cybernetics. The Technology Press, New-York etc. 1949, 194 pp.

*Учебное издание*

***Марчук** Юрий Николаевич*

# КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

Корректор *Т. И. Лошкарева*
Верстка и оформление *Е. В. Романова*

«Восток – Запад»

Тел./факс: (495) 101-36-29
Для корреспонденции: 127106, Москва, а/я 12
E-mail: info@muravei. ru
Интернет: www. muravei. ru

Общероссийский классификатор продукции
ОК-005-93, том 2; 953005 — учебная литература

Санитарно-эпидемиологическое заключение
№ 77.99.02.953.Д.003857.05.06 от 05.05.2006 г.

ООО «Издательство АСТ»
170002, Россия, г. Тверь, пр. Чайковского, д. 27/32
Наши электронные адреса:
WWW.AST.RU E-mail: astpub@aha.ru

ООО «Восток — Запад»
129085, г. Москва, Звездный бульвар, 21, стр. 1

Отпечатано с готового оригинал-макета в
ООО «Типография ИПО профсоюзов Профиздат»,
109044, Москва, Крутицкий вал, 18.