ВОПРОСЫ ЯЗЫКОЗНАНИЯ

№ 5 1990

БЕЛОНОГОВ Г. Г., КУЗНЕЦОВ Б. А., НОВОСЕЛОВ А. П.,
ПАЩЕНКО Н. А.

ЛИНГВИСТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ АВТОМАТИЗИРОВАННЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

Последние два десятилетия характеризуются широким внедрением электронной вычислительной техники в различные области человеческой деятельности: науку, технику, экономику, военное дело и др. При этом ЭВМ стали использоваться не только и не столько для автоматизации вычислений, а прежде всего для решения информационных задач. Это привело к созданию автоматизированных информационных систем.

Автоматизированные информационные системы (АИС) предназначены для автоматизации процессов накопления, поиска и обработки информации. При их создании возникает широкий круг проблем, среди которых важное место занимают проблемы технического, программного, информационного и лингвистического обеспечения, тесно связанные между собой. Так, информационное и лингвистическое обеспечение АИС создается с учетом возможностей технических средств. То же самое можно сказать и о программном обеспечении. С другой стороны, вновь создаваемые технические средства должны быть ориентированы на перспективные методы автоматической обработки информации.

Наиболее трудными для решения являются обозначенные ниже проблемы, связанные с содержательной, смысловой стороной информационной технологии.

1. Проблема адекватного представления знаний и вытекающая из нее — разработки эффективных формализованных информационных языков. Сюда же относится и проблема формализованного представления смысла текстов, вводимых в ЭВМ на естественных языках.

2. Проблема автоматизации преобразования информации из одной формы представления в другую, например, ее перевода с естественного языка на формализованный информационный язык, с одного естественного языка на другой или с одного формализованного языка на другой.

3. Проблема автоматического распознавания смыслового тождества высказываний, представленных в различной языковой форме.

4. Проблема автоматизированного получения данных на основе информации, хранящейся в АИС.

5. Проблема общения человека с АИС.

Большинство перечисленных проблем требует для своего решения привлечения наряду с другими методами также и лингвистических методов. Тем не менее на первых этапах развития АИС лингвистические вопросы, связанные с их созданием, разрабатывались не только и не столько лингвистами, сколько специалистами другого профиля (математиками, инженерами). Постепенно сформировалась область деятельности, которая в конце 60-х годов получила название «лингвистическое обеспечение».

В настоящее время под лингвистическим обеспечением АИС обычно понимают комплекс мероприятий, связанных с разработкой, ведением, и применением лингвистических средств, а также сами эти средства. Лингвистические средства АИС включают собственно языковые средства н процедуры (программы) обработки текстовой информации. Языковыми средствами АИС являются естественный язык и его производные — ограниченный естественный язык и формализованные языки. Процедурными — средства манипулирования смысловыми единицами различных уровней (морфемами, словами, словосочетаниями, фразами, сверхфразовымн единствами), и в частности средства их семантико-синтаксического анализа и синтеза.

В современных АИС формализованное представление информации базируется на концепциях математической логики, и прежде всего того ее раздела, который носит название «исчисление предикатов». Это и не удивительно, так как ЭВМ ориентированы на манипулирование предикатно-актантными структурами данных. К такому выводу легко прийти, изучая структуру машинных операций и структуру алгоритмических языков высокого уровня (таких, как Алгол, Кобол, Фортран, PL/I, Лисп и др.).

Предикатно-актантная структура данных строится на основе многоместных предикатов, которые имеют вид

$$F(\ ,\ ,\ \ldots\ ,) \cdot \qquad (1)$$

Здесь F — имя предиката (многоместного отношения), а пустые места предназначены для актантов (значений предметных переменных). Конкретные высказывания формируются путем подстановки на пустые места значений предметных переменных, соответствующих описываемым ситуациям, процессам или объектам. Так, высказывание о ситуации, в которой выделено п элементов, будет иметь вид:

$$F\ (X_x,\ X_2.\ \ldots\ .,\ X_п) \qquad (2)$$

Здесь F, как и ранее, — имя понятия, обозначающего предикат, Хц $X_2\ \ldots\ ., X_п$ — имена понятий, обозначающих элементы, входящие в состав ситуации. От структуры (2) можно легко перейти к структуре в виде конкатенации (связки, сочетания) двусоставных признаков, состоящих из их наименований и значений. Значения признаков будут представлены именами элементов в высказывании.

В АИС отображаются явления внешнего «мира» (внешнего по отношению к АИС), и в качестве элементов этого «мира» выступают его объекты (материальные или абстрактные). Членение внешнего «мира» на объекты может быть разным и зависит от целевой установки. Объекты могут быть простыми и сложными. Простой объект воспринимается как носитель совокупности характеризующих его свойств. Внутренняя структура простого объекта не раскрывается. Сложный объект состоит из простых объектов (как минимум двух), связанных между собой. Он также воспринимается как нечто целое и характеризуется определенными свойствами. Но в отличие от простого объекта в нем различается внутренняя структура — его расчлененность на простые объекты. Деление объектов на простые и сложные относительно: один и тот же объект внешнего мира может при решении одних задач рассматриваться как простой, а при решении других — как сложный.

Свойствам объектов в информационном отображении соответствуют их признаки, но в АИС отображаются не все свойства объектов, а лишь наиболее существенные, причем оценка существенности тех или иных свойстн

зависит от характера решаемых задач. Простому объекту внешнего мира в информационном отображении соответствует конкатенация характеризующих его признаков, а сложному — сетевая структура. В узлах этой структуры помещаются простые объекты, а узлы соединяются дугами, которые отражают связи (бинарные отношения) между объектами.

Понятия «бинарное отношение» и «признак» во многом сходны друг с другом. И то, и другое характеризует определенное свойство объекта: первое — «находиться в определенном отношении к другому объекту», второе — «соотноситься с определенной качественной или количественной категорией». Более того, бинарное отношение можно считать частным случаем признака, характеризующим связь объекта с некоторым другим объектом. Частным случаем признака можно считать и математическое понятие переменной: наименование переменной может быть интерпретировано как наименование признака, а значение переменной — как значение признака.

При описании объектов на формализованных информационных языках в качестве минимальной самостоятельной единицы смысла выступает элементарное высказывание, в котором утверждается принадлежность объекту одного из его признаков. Признак может выражаться одним понятием, но обычно он расчленяется на две части: на наименование признака и его значение. Таким образом, элементарное высказывание может быть представлено в виде триады, состоящей из идентификатора объекта, наименования признака и его значения. Все элементы этой триады присутствуют во всех формализованных языках, но кодируются они по-разному: часть элементов кодируется позиционными средствами, другая — комбинациями символов алфавита. В соответствии с этим в АИС применяются три основных формата высказываний — позиционный, анкетный и триадный, которые могут использоваться самостоятельно и в различных сочетаниях. В позиционном формате для каждого признака отводится определенное поле памяти, на котором записываются значения этого признака (наименования признаков представляются позициями соответствующих им значений и в явном виде не записываются). Связь между признаками обозначается контактным расположением полей, предназначенных для описания одного объекта. В анкетном формате наименования и значения признаков обозначаются в явном виде — комбинациями символов алфавита, а связь между признаками — их контактным расположением. Порядок следования признаков в пределах одного высказывания не имеет значения. В триадном формате все компоненты элементарных высказываний — идентификаторы объектов, наименования признаков и их значения — выражаются комбинациями символов алфавита.

Популярной формой представления формализованной информации в позиционном формате являются двумерные таблицы. В таких таблицах в качестве наименований граф используются обобщенные наименования объектов и наимзнования признаков объектов. В графах записываются наименования конкретных объектов и соответствующие этим объектам значения признаков (числовые или текстовые).

Подводя итогп сказанному, можно утверждать, что наиболее общим видом формализованного описания сложных объектов является сетевая структура, в узлах которой помещены описания простых объектов, а узлы соединены друг с другом дугами, обозначающими отношения между простыми объектами. Такая структура может быть изображена в виде линейной последовательности описаний всех ее узлов, где каждый узел в свою очередь представляется в виде конкатенации собственных признаков обоз-

начаемого им ооъекта и признаков связи этого объекта с другими объектами. Конкатенация признаков может быть оформлена в виде позиционной, анкетной или триадной структуры или в виде сочетания таких структур. Частным случаем сетевой структуры является иерархическая структура (например, дерево зависимостей, отображающее синтаксическую структуру предложения).

Выше мы говорили о том. что в основе формализованного представления информации в памяти ЭВМ лежит предикатно-актантиая структура. Эта структура используется также при описании единиц и структур естественных языков. Различные ее вариации применяются для представления семантических множителей, семантических падежей, семантических сетей, концептуальных сетей, фреймов и т. д.

АИС принято делить на документальные и фактографические. В документальных системах хранятся и обрабатываются обобщенные сведения о научно-технических документах (библиографические описания, рефераты) или тексты документов, в фактографических системах — сведения о признаках объектов любой другой природы (о свойствах веществ и материалов, о характеристиках промышленных изделий, о производственных показателях, о запасах сырья, о состоянии окружающей среды и т. п.).

Основными средствами лингвистического обеспечения документальных систем являются: 1) рубрикаторы и классификаторы информации (рубрикатор Государственной автоматизированной системы научно-технической информации, Международная классификация изобретений, Универсальная десятичная классификация, отраслевые рубрикаторы и классификаторы); 2) тезаурусы и другие нормативные словари (в настоящее время различными организациями СССР создано более 120 тезаурусов); 3) унифицированные коммуникативные форматы представления информации на машиночитаемых носителях; 4) программные средства перевода информации из одного ее представления в другое — конверторы; 5) базовые процедуры для автоматической обработки текстов на естественном языке (процедуры морфологического п семантико-синтаксического анализа и синтеза): 6) машинные словари слов и терминологических словосочетаний, а также грамматические таблицы для автоматической обработки текстов; 7) процедуры автоматического составления и лингвистической обработки машинных словарей; 8) процедуры автоматизированного обнаружения и исправления ошибок в текстах документов при их вводе в ЭВМ; 9) процедуры автоматического индексирования документов и запросов (их перевода с естественного языка на формализованный информационный язык); 10) процедуры автоматической классификации документов; 11) нормативно-техническая документация по лингвистическому обеспечению документальных АИС (Положение о лингвистическом обеспечении Государственной автоматизированной системы научно-технической информации. Государственные стандарты по лингвистическому обеспечению, инструкции и др.).

Поиск научно-технических документов в документальных АИС обычно производится по их формализованным описаниям или по текстам рефератов. В последнее время стали также создаваться системы, в которых хранятся полные тексты документов, и поиск ведется по этим текстам. В процессе поиска допускается неполная выдача документов, удовлетворяющих условиям запросов («потери»), и выдача лишних документов («шумы»). В фактографических системах это исключено. Поиск без потерь и шумов здесь обеспечивается за счет формализации информации и строгого контроля за использованием языковых средств. Формализация и контроль

осуществляются на основе применения табличных или анкетных форм ввода информации в ЭВМ.

Каждая форма ввода информации создается для описания одного класса однородных объектов и определяется регламентированным перечнем наименований признаков, а для каждого признака указывается перечень допустимых значений. Для числовых характеристик указываются единицы измерения. Вместо перечней допустимых значений нечисловых характеристик может даваться ссылка на рекомендуемые для использования классификаторы.

Для облегчения процессов формализации информации средства лингвистического обеспечения (перечни форм документов и допустимых значений признаков) могут храниться в памяти ЭВМ, а манипулирование ими при вводе информации и при формулировке поисковых предписаний может производиться в диалоговом режиме (в режиме «меню», пли подсказки). В развитых АИС наряду с процедурными средствами ввода, обновления, поиска и редактирования информации должна быть библиотека прикладных программ для решения расчетных задач. Если в такого рода системах имеются интерфейсы для перехода от базового представления информации к ее представлению, необходимому на входе прикладных программ, и от представления на выходе прикладных программ снова к базовому представлению, то то1да сравнительно легко могут быть реализованы сложные комплексы расчетных задач.

Важной разновидностью фактографических АИС являются экспертные системы. В [1] экспертные системы определяются как интеллектуальные программы, использующие формализованные знания и процедуры логического вывода для решения проблем, которые достаточно трудны и требуют большого опыта. Знания, закладываемые в экспертные системы, представляют собой совокупность фактов и эвристик. Эвристики — это правила рассуждений, которыми руководствуется квалифицированный специалист при принятии решений в той или иной предметной области. В общем случае экспертная система должна включать базу знаний, процедуры логического вывода и интерфейс, позволяющий неискушенному пользователю общаться с этой системой на естественном языке. В [1] отмечается, что в настоящее время наименее развитой частью экспертных систем является естественноязыковой интерфейс.

Экспертные системы находят широкое применение в различных областях человеческой деятельности — медицинская диагностика, анализ и интерпретация экспериментальных данных, проектирование, планирование, обучение, распознавание образов и др. В них используются разные способы представления знаний. Наиболее популярным из них является система продукций — система правил типа «ситуация — действие». База знаний экспертной системы может включать сотни и даже тысячи таких правил. В [Г] отмечается, что система продукций как способ представления знаний не всегда оказывается надежной. Автор статьи полагает, что в будущем экспертные системы должны базироваться на способах представления знаний, позволяющих более адекватно отражать структуры объектов и причинно-следственные отношения (на семантических сетях, фреймах и др.).

Естественноязыковой интерфейс в общем случае должен базироваться на процедурах семантико-синтаксического анализа и синтеза текстов и необходимых для этой цели системах словарей и грамматических таблиц. Можно, например, представить себе процесс анализа текстов (сообщений пли запросов) на входе АИС, состоящий из процедур морфологического, синтаксического и концептуального анализа.

В процессе морфологического анализа производится членение слов на морфемы или сочетания морфем, отождествление их с соответствующими элементами словаря и определение грамматических характеристик слов (части речи, род, число, падеж, лицо, модели управления и др.), необходимых на последующих этапах анализа. Грамматические характеристики определяются с той степенью точности, которая возможна на основе анализа буквенного состава словоформ без учета их контекстного окружения.

В процессе синтаксического анализа строится формализованная модель текста (в общем случае сетевая), в узлах которой находятся слова, а дуги отражают синтагматические отношения между словами (чаще всего отношения непосредственной доминации типа «определяющее — определяемое» или «управляющее —управляемое»). Формализованная модель текста может строиться в виде последовательности формализованных моделей составляющих его предложений или (что лучше) — учитывать и межфазовые связи.

В процессе концептуального анализа происходит распознавание наименований понятий и отношений между ними на основе сопоставления формализованной модели текста и наименований понятий, представленных в АИС. При этом могут устанавливаться не только отношения между понятиями, имена которых содержатся непосредственно в тексте (синтагматические отношения), но и отношения понятий из текста с другими понятиями (парадигматические отношения).

Системы словарей и грамматических таблиц, необходимые для семантико-синтаксического анализа текстов, могут быть различными в зависимости от принятой концепции такого анализа. Например, при морфологическом анализе — словарь основ слов, словарь окончаний, словарь суффиксов (сочетаний суффиксов) и грамматические таблицы, характеризующие системы словоизменения и словообразования. При синтаксическом анализе — словари синтагм, представленные сочетаниями символов классов слов. При концептуальном анализе — словари наименований понятий (словосочетаний и слов) и словари парадигматических отношений между понятиями (тезаурусы, концептуальные сети).

Словари и грамматические таблицы, применяемые при семантико-синтаксическом анализе текстов, используются, как правило, и при их семантико-синтаксическом синтезе. На первом этапе синтеза (назовем его этапом концептуального синтеза) числовые коды понятий в сообщениях, выдаваемых из АИС, заменяются с помощью соответствующего словаря на наименования понятий (словосочетания или слова). Далее (на этапе синтаксического синтеза) вырабатывается грамматическая информация, необходимая для оформления предложений на естественном языке (информация о синтаксической структуре предложений и о формах входящих в них слов). Наконец (на этапе морфологического синтеза), словам текста придается форма, обусловленная их ролью в составе предложений.

Следует заметить, что в системах автоматической обработки информации «смысл» текста нельзя распознать, опираясь только на текст. Как известно, не все в нем выражается в явном виде, многое подразумевается. Осложняют анализ текстов и такие явления, как полисемия лексических единиц, эллипсис, наличие анафорических связей между предложениями, метафория и др. Поэтому для распознавания «смысла» сообщений, вводимых в АИС. необходимо наряду с текстами этих сообщений, словарями и грамматическими таблицами привлекать также «экстралингвистическую» базу знаний (базу знаний АИС), и процедуры, моделирующие процессы мышления (например, процедуры логического вывода). Корректность

распознавания «смысла» сообщений зависит не только от качества используемых лингвистических моделей анализа текстов, но и от качества баз знаний, хранящихся в АИС, а также от качества применяемых моделей человеческого мышления.

За последние 15—20 лет в СССР и за рубежом сделано немало для теоретического осмысления процессов автоматического анализа и синтеза текстов и их практической реализации. Этой проблеме посвящены десятки международных и национальных научных конференций, сотни монографий и тысячи статей [2—6]. Однако следует признать, что для ее решения необходим широкий фронт исследований, базирующийся на применении методов различных наук, и прежде всего на сочетании традиционных лингвистических методов и методов, используемых при создании АИС (в том числе методов компьютерной лингвистики и методов моделирования процессов мышления, разрабатываемых специалистами по «искусственному интеллекту»).

В последнее время возникли большие надежды на образование такого фронта в рамках программы Машинного фонда русского языка [7, 8], инициаторами создания которого были А. И. Ершов и Ю. Н. Караулов. Идея была поддержана рядом других ученых.

Наиболее актуальной практической задачей лингвистического обеспечения АИС является создание машинных словарей (словарей слов, словарей словосочетаний, тезаурусов и др.) и базовых процедур автоматической обработки текстов (например, процедур морфологического и синтаксического анализа и синтеза). На основе этих словарей и процедур могут создаваться различные комплексные процедуры — процедура обнаружения и исправления ошибок в текстах при их вводе в ЭВМ, процедура перевода текстов с естественного языка на формализованный информационный язык, процедура перевода текстов с одного естественного языка на другой и т. п.

Практические задачи лингвистического обеспечения АИС решаются во многих организациях Советского Союза, в том числе (и не в последнюю очередь) в центрах научно-технической информации (таких, как ВИНИТИ, ВНТИЦентр, ИНИОН, ВНИИКИ, Информэлектро и др.). Характер работ, проводимых в этих центрах, мы проиллюстрируем на примере ВИНИТИ.

В ВИНИТИ АН СССР работы по лингвистическому обеспечению АИС ведутся как в направлении создания словарных, так и процедурных средств. II словарные, и процедурные средства разрабатываются с целью повышения качества документальных баз данных на машиночитаемых носителях и качества поиска в них. Научно-технические документы (статьи из журналов, книги, доклады на конференциях, патенты и др.) представляются в базах данных в виде библиографических описаний, сопровождаемых поисковыми образами этих документов, а в значительной части — не менее 30% — и текстами рефератов. Ежегодно на машиночитаемые носители переносится более одного миллиона описаний документов-публикуемых в нашей стране и за рубежом. Не учитываются лишь публикации по сельскому хозяйству, медицине, строительству и архитектуре (они обрабатываются в других центрах научно-технической информации).

Базы данных на машиночитаемых носителях являются хорошим исходным материалом для составления словарей, и эта возможность широко используется в ВИНИТИ. В настоящее время там составлен политематический машинный словарь словоизменительных основ слов объемом более 180 тыс. лексических единиц. Каждая словоизменительная основа сопро-

вождается информацией о ее длине, о длине входящей в ее состав словообразовательной основы, о принадлежности к части речи, о типе словоизменения, о модели управления. Словарь составлялся по текстам протяженностью более 70 млн. слов. При этом обрабатывались тексты по машиностроению, электротехнике, энергетике, автоматике, радиоэлектронике, информатике, вычислительной технике, механике, транспорту, геологии, горному делу, металлургии, физике, охране окружающей среды и др. Словарь покрывает тексты любой тематики по естественным и техническим наукам на 98—09%. В этот словарь был влит машинный словарь основ слов по химико-биологической тематике, который имел первоначальный объем более 100 тыс. основ слов и обеспечивал лексическое покрытие текстов по биологии и химии на 97—98%.

В ВИНИТИ проводится также работа по составлению частотных словарей терминологических словосочетаний. Исходным материалом для словарей служат поисковые образы документов на машиночитаемых носителях, в которых указаны перечни слов и словосочетаний, характеризующие тематическое содержание документов. В течение 1985—1989 гг. было составлено 15 частотных словарей, охватывающих все тематические области, представленные в базах данных ВИНИТИ. При этом были обработаны поисковые образы документов общей протяженностью более 17 млн. слов. Суммарный объем словарей составил более 300 тыс. терминов, встретившихся в поисковых образах документов два и более раз.

При составлении машинных словарей основ по текстам наиболее трудоемкой операцией является лингвистическая обработка словоформ (выделение у них словоизменительных и словообразовательных основ, определение принадлежности к части речи, типа словоизменения и т. п.). Выполнение этой операции может быть в значительной мере автоматизировано. Возможность автоматизации определяется имеющей место связью между буквенным составом слов, с одной стороны, и их морфологической структурой и грамматическими признаками, с другой. Например, словоформы с одинаковыми буквенными кодами своих концов обладают, как правило, и одинаковым морфемным составом этих концов (состав суффиксов и окончаний), а также совпадающими грамматическими признаками. Это имеет место и в тех случаях, когда словоизменительные и словообразовательные основы у словоформ различны.

Лингвистическая обработка словоформ текста может осуществляться с помощью эталонного обратного словаря словоформ, в котором вручную заранее были выделены словоизменительные и словообразовательные основы и указаны грамматические признаки. В процессе обработки для каждой словоформы ищется максимальное покрытие справа одной из словоформ эталонного словаря, и у нее вычленяются окончание и словообразовательные суффиксы, совпадающие с окончанием и словообразовательными суффиксами словарной словоформы. Таким образом выделяются словоизменительная и словообразовательная основы. Затем грамматические признаки словарной словоформы переносятся на анализируемую словоформу.

Алгоритм лингвистической обработки словарей реализован на ЭВМ ЕС-1055 и работает при объеме эталонного словаря около 45 тыс. лексических единиц со скоростью более 200 слов в секунду [1]. При этом словоизменительные основы у «новых» слов выделяются правильно с вероят-

[1] В настоящее время программы лингвистической обработки словарей, морфологического анализа текста и нормализации слов переведены на персональные ЭВМ типа IBM PC XT/AT.

ностью 0.99, словообразовательные основы — с вероятностью 0,98, тип словоизменения определяется правильно с вероятностью 0,90, а части речи — с вероятностью 0..98. Результаты работы алгоритма выдаются на печать в удобном для восприятия виде, проверяются человеком и при необходимости корректируются. Корректирующая информация вводится в ЭВМ и с помощью специальной программы вносится в обрабатываемый машинный словарь. Такой порядок работы позволяет в значительной мере (в десятки раз) сократить затраты труда на создание машинных словарей, т. к. все признаки «новых» слов генерируются автоматически, а корректирующая информация составляет лишь небольшую долю общего объема этих словарей.

Машинные словари основ слов (политематический и химико-биологический) используются в программе морфологического анализа [91. Эта программа на ЭВМ ЕС-1055 может обрабатывать тексты со скоростью более 300 слов в секунду. Такая высокая скорость достигается за счет применения специальной организации машинного словаря (так называемой ассоциативно-адресной структуры) и методов хеширования, позволяющих вычислять адреса хранения основ слов в словаре по их буквенным кодам.

В процессе морфологического анализа для словоформ текста определяются длина словоизменительной и словообразовательной основ, часть речи, номер флективного класса (номер списка окончаний, совместимых со словоизменительной основой), модель управления, а также наборы таких характеристик, как род, число, падеж и лицо. Перечисленные признаки определяются не только для тех словоформ, основы которых включены в машинный словарь, но и для любых других словоформ. Это оказалось возможным благодаря применению принципа грамматической аналогии, о котором речь шла выше при рассмотрении процедуры автоматизации лингвистической обработки словарей.

Программа морфологического анализа используется в ряде процедур автоматической обработки текстов и словарей. Одной из таких процедур является процедура автоматической нормализации (лемматизации) слов. Процедура нормализации позволяет преобразовывать текстовые формы слов в некоторые приоритетные формы, принятые за канонические. Например. существительные, за исключением pluralia tantnm, приводятся к форме им. падежа ед. числа (pluralia tanlum — к форме им. падежа мн. числа): прилагательные — к форме им. падежа ед. числа муж. рода; глаголы, полные и краткие причастия и деепричастия — к форме инфинитива. Форма наречий, союзов, предлогов и частиц оставляется без изменений.

В основу построения процедуры автоматической нормализации слов также положен принцип аналогии, который применительно к рассматриваемой задаче можно сформулировать следующим образом: если у каких-либо слов совпадает буквенный состав их концов, то они, как правило, имеют и одинаковые модели словоизменения и словообразования (при этом совпадающие концы слов не обязательно должны состоять только из суффиксов и окончаний).

Правила перехода от текстовых форм слов к каноническим строятся по формуле «ситуация — действие». В качестве идентификаторов ситуаций здесь используются флективные классы слов и конечные буквосочетания их основ (иногда также признак «глагольности»), а выполняемые действия заключаются в отделении от основ слов заданного числа букв и добавления к ним заданных нормализующих буквосочетаний. Например, продукция, позволяющая преобразовать словоформу *накрашенного*

в каноническую форму *накрасить,* будет иметь в качестве идентификатора ситуации флективный класс 103 (класс прилагательных, склоняющихся как слово *главный)* и конечное буквосочетание основы *-ашенн-;* действие по нормализации будет заключаться в отделении от конца основы *-накрашенн-* четырех букв и присоединении к ней буквосочетания *-ить.* Аналогично переход от словоформы *зубца* к канонической форме *зубец* позволит сделать продукция, у которой идентификатором ситуации является флективный класс 001 (класс слов, склоняющихся как слово *стол)* и буквосочетание *бц,* а действие по нормализациии будет состоять в отделении от основы слова одной буквы и добавлении к ней нормализующего буквосочетания *-ец.* Таким же способом могут быть выполнены и преобразования типа *колец — кольцо, огня — огонь, введенный — ввести, ведя — вести, могут — мочь, ищут — искать, прошел — пройти.*

Список продукций для автоматической нормализации слов, составленный на основе анализа обратного политематического словаря основ, включает более 700 правил. Некоторые из этих правил могут быть применены только к одному слову, но большинство — к целым классам слов. Нормализация слов, отсутствующих в машинном словаре, ничем не отличается от нормализации слов, содержащихся в нем, т. к. необходимые для этого исходные данные (номер флективного класса, длина словоизменительной основы и др.) получаются в процессе морфологического анализа. Скорость нормализации на ЭВМ типа ЕС-1055 составляет более 200 слов в секунду, а если не учитывать время, необходимое для морфологического анализа, то около 1000 слов в секунду.

Процедуры морфологического анализа и нормализации слов позволили решить ряд других прикладных задач лингвистического обеспечения АНС, например, задачу автоматического орфографического контроля текстов при их вводе в ЭВМ, задачу автоматического индексирования документов для их последующего поиска, задачу автоматического составления словарей словосочетаний по неформализованным текстам и др.

Орфографический контроль текстов строится как процесс их морфологического анализа и выдачи неопознанных (отсутствующих в машинном словаре) слов для опознания человеком. Не опознанные в процессе морфологического анализа слова могут выдаваться на экран дисплея или на печатающее устройство. Они сопровождаются микроконтекстом. Человек, опираясь на микроконтекст, отмечает искаженные слова и исправляет их. На ЭВМ ЕС-1033 просмотр текстов с целью их орфографического контроля ведется со скоростью более 400 слов в секунду (около 300 авторских листов в час).

Процедура автоматического индексирования документов реализована в ВИНИТИ в двух вариантах — в виде системы пословной нормализации текстов (промышленный вариант) и в виде системы автоматического составления поисковых образов документов (экспериментальный вариант). В первом случае обеспечивается повышение качества и комфортность поиска документов (человек при формулировке запросов оперирует не многообразием форм слов русского языка, а нормализованными формами). Во втором — исключается трудоемкая ручная операция по составлению поисковых образов документов и появляется возможность автоматического составления указателей к реферативным журналам.

Процедура автоматического составления словарей словосочетаний по неформализованным текстам (например, по заголовкам документов и текстам их рефератов) применяется как дополнительное средство выявления терминологической лексики (наряду с обработкой поисковых образов до-

кументов). Эта процедура выполняется за четыре этапа: 1) морфологический анализ текстов; 2) выделение именных словосочетаний из текстов; 3) составление частотного словаря словосочетаний; 4) нормализация составленного словаря.

Первый этап нами уже описан. Второй представляет собой по сути процедуру синтаксического анализа текста. Он был реализован в двух вариантах — в виде процедуры, базирующейся на построении дерева зависимостей, и в виде процедуры локального синтаксического анализа. Второй вариант оказался более удачным. Именные словосочетания здесь выделялись как грамматически согласованные непрерывные последовательности существительных и прилагательных. Знаки препинания и все части речи, кроме существительных и прилагательных, рассматривались как границы именных словосочетаний.

На третьем этапе составления частотного словаря по выделенным словосочетаниям строились их поисковые образы — последовательности основ слов, в которых слева стояли словоизменительные основы опорных слов словосочетаний, а за ними в алфавитном порядке располагались словообразовательные основы их определителей. Далее массив выделенных словосочетаний сортировался по алфавиту поисковых образов и из каждой группы словосочетаний с одинаковыми поисковыми образами в частотный словарь включалась только одна форма словосочетания, которая сопровождалась частотой встречаемости ее поискового образа. После этого из словаря исключались малочастотные словосочетания.

На четвертом этапе составления словаря проводилась нормализация включенных в него текстовых форм именных словосочетаний. При этом опорному слову каждого словосочетания (первому слева существительному) придавалась форма им. падежа ед. или мн. числа, а определяющие его прилагательные согласовывались с ним в роде, числе и падеже. В заключение проводилось редактирование составленного словаря (из него исключались неинформативные словосочетания). Опыт составления словарей подтвердил эффективность описанной процедуры: словарь, составленный по текстам рефератов по информатике и вычислительной технике объемом в один миллион триста тысяч слов, оказался вполне удовлетворительным, а в процессе редактирования из него пришлось исключить не более 15% неинформативных словосочетаний.

Возможность составления словарей словосочетаний по поисковым образам документов и по текстам их рефератов создает предпосылки для автоматизированного составления тезаурусов по различным областям науки и техники. В ВИНИТИ проводятся довольно успешные эксперименты в этом направлении. В экспериментах парадигматические связи между словосочетаниями устанавливаются на основе информации о парадигматических связях между входящими в их состав словами. Далее результаты машинных решений корректируются человеком.

В заключение следует сказать, что несмотря на несомненные успехи в деле лингвистического обеспечения автоматизированных информационных систем главные трудности здесь еще впереди. Они будут связаны прежде всего с разработкой адекватных процедур семантико-синтаксического анализа и синтеза текстов и с созданием мощных машинных словарей, содержащих достаточно богатую информацию о синтагматических и парадигматических признаках входящих в их состав единиц языка и речи.

СПИСОК ЛИТЕРАТУРЫ

1. *Gevarter W. B.* Expert systems : artificial intelligence applied // Telematics and informatics. 1984. V. 1. № 3.
2. *Виноград Т.* Программа, понимающая естественный язык. М., 1976.
3. *Шенк Р.* Обработка концептуальной информации. М., 1980.
4. *Караулов Ю.Н.* Лингвистическое конструирование и тезаурус литературного языка. М., 1981.
5. *Кулагина О. С.* Исследования по машинному переводу. М., 1979.
6. *Попов Э. В.* Общение с ЭВМ на естественном языке. М., 1982.
7. *Ершов А. П.* К методологии диалоговых систем. Феномен деловой прозы *II* Вопросы кибернетики. Общение с ЭВМ на естественном языке / Под ред. Розенцвенга В. Ю . М., 1982.
8. *Андрющенко В. М.* Машинный фонд русского языка: идеи и суждения // Вопросы кибернетики. Прикладные аспекты лингвистической теории / Под ред. Ершова А. П. М., 1987.
9. *Белоногое Г. Г., Зеленков 10. Г.* Алгоритм морфологического анализа русских слов // Вопросы информационной теории и практики. Автоматизированная словарная служба. Автоматическое индексировние документов/ Под ред. Белоногова Г. Г., М., 1985.