РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ ПРОБЛЕМ ПЕРЕДАЧИ ИНФОРМАЦИИ

ЛИНГВИСТИЧЕСКИЙ ПРОЦЕССОР ДЛЯ СЛОЖНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

Ответственный редактор
доктор филологических наук
Л. П. Крысин

МОСКВА "НАУКА"
1992

Авторы:

Ю.Д. АПРЕСЯН, И.М. БОГУСЛАВСКИЙ, Л.Л. ИОМДИН,
А.В. ЛАЗУРСКИЙ, Л.Г. МИТЮШИН, В.З. САННИКОВ, Л.Л. ЦИНМАН

УДК 519.766

Лингвистический процессор для сложных информационных систем / Ю.Д. Апресян, И.М. Богуславский, Л.Л. Иомдин и др. – М.: Наука, 1992, – 256 с. – ISBN 5–02–006928–0.

Лингвистическим процессором называется реализованная на ЭВМ формальная лингвистическая модель, способная понимать и производить тексты на неограниченном языке. Она включает три основных массива правил – морфологические, синтаксические и семантические – и обслуживающие их словари. Эти компоненты обеспечивают пофразное преобразование текста в морфологические, синтаксические и семантические структуры и обратно. Лингвистический процессор был экспериментально опробован в подсистеме общения с базами данных на естественном языке, а также в системах машинного перевода с английского языка на русский и с русского на английский.

Для специалистов в области вычислительной лингвистики, информатики, общего языкознания и лексикографии.

Библиогр.: 65 назв.

A linguistic processor for complex information sistems / Yu.D. Apresjan, I.M. Boguslavskij, L.L. Iomdin and all. – Moscow: Nauka, 1992. – 256 p.

A linguistic processor is a computer–implemented formal linguistic model which is able to understand and produce texts written in free natural language. The model, presented in the book, includes three major sets of rules: morphological, syntactic, and semantic rules, as well as dictionaries with which these sets of rules interact. These components provide for sentence–by–sentence conversion of natural language texts into morphological, syntactic, and semantic structures, and vice versa. The linguistic processor was tested experimentally in a natural language interface system, intended to facilitate man–computer communication, as well as in an English–to–Russian and a Russian–to–English machine translation systems.

The book is intended for specialists and students in computational linguistics, information science, general linguistics, and lexicography.

Рецензенты:

В.А. УСПЕНСКИЙ, В.Ю. РОЗЕНЦВЕЙГ

Л $\frac{4602010000-020}{042\,(2)-92}$ 513-92 II полугодие

ISBN 5–02–006928–0

Глава 1
ОБЩЕЕ ПРЕДСТАВЛЕНИЕ О ЛИНГВИСТИЧЕСКОМ ПРОЦЕССОРЕ: СТРУКТУРА, НАЗНАЧЕНИЕ И ПРИНЦИПЫ РАЗРАБОТКИ

1.1. Постановка задачи

Попытки формализовать интеллектуальную деятельность человека привели к постановке фундаментальной лингвистической задачи, состоящей в моделировании его **языкового поведения**, т.е. в построении функциональной кибернетической модели естественного языка (ЕЯ). Естественный язык служит человеку для выражения собственных мыслей и для понимания мыслей других людей. С известным огрублением можно сказать, что первому виду языковой деятельности соответствует производство текстов на ЕЯ, а второму - понимание таких текстов. Если обозначить множество текстов через {T}, а множество выражаемых ими смыслов через {C}, то модель ЕЯ можно определить как транслятор, устанавливающий соответствие (1) между этими двумя множествами:

(1) $$\{T\} \leftrightarrow \{C\}$$

Формальные модели языка, разрабатывавшиеся первоначально в чисто теоретическом плане (см., напр.: [Chomsky, 1956; Мельчук, 1974]), в последнее время все чаще рассматриваются как компоненты различных прикладных систем. Будучи реализованы на компьютере, они входят в качестве составных частей в системы машинного перевода (МП), подсистемы общения с базами данных (БД) на неограниченном ЕЯ и другие сложные информационные системы. Дж. Симонс считает разработку систем, способных понимать ЕЯ, "основной целью исследований в области искусственного интеллекта" [Симонс, 1985; с. 125].

Будем называть компьютерную систему, реализующую формальную лингвистическую модель и способную работать с ЕЯ во всем его объеме, лингвистическим процессором (ЛП). В современной информатике лингвистическими процессорами называются и другие средства переработки текстовой информации на ЕЯ, в том числе и не рассчитанные на работу с ЕЯ в полном объеме. Однако мы используем термин "лингвистический процессор"

лишь в указанном, более узком понимании, а в настоящей монографии применяем его лишь к разрабатываемой нами системе.

Как следует из сказанного выше, две основные функции ЛП состоят в извлечении смысла из заданного текста (моделирование понимания, анализ) и в выражении заданного смысла текстом на ЕЯ (моделирование производства текстов, синтез).

Обычно говорят о понимании текстов в слабом и сильном смысле. Понимание в слабом смысле имеет место тогда, когда обрабатываемый текст может быть перифразирован средствами того же самого или другого языка. Эта модель понимания реализуется, например, при переводе текста с одного ЕЯ на другой: перевод есть пересказ содержания текста, написанного на одном языке, средствами другого языка. Понимание в сильном смысле имеет место тогда, когда воспринимаемый текст требует от адресата определенной реакции, и когда эта реакция действительно имеет место. Эта модель понимания реализуется, например, при общении с БД, когда последняя выдает правильный ответ на обращенный к ней вопрос.

Для моделирования понимания в обоих указанных смыслах в любых системах, претендующих на глубокую переработку естественно-языковых текстов, необходимо иметь особый уровень представления высказываний, который можно назвать семантическим. Он задается формальным семантическим языком, выразительные средства которого достаточно велики для того, чтобы отразить полностью содержание текста на исходном ЕЯ в рамках поставленной задачи.

Располагая таким языком, можно формально описать процесс понимания (анализа) текстов как перевод с естественного языка на семантический, а процесс производства текстов (синтез) - как "обратный" перевод с семантического языка на естественный. О том, как понимается формальный семантический язык, мы скажем ниже. Что касается текста на ЕЯ, то под ним понимается последовательность предложений в обычной орфографической записи. Не предполагается осуществлять распознавание и порождение звучащей речи, хотя в принципе ЛП может быть дополнен соответствующими фонетическими компонентами с тем, чтобы на вход анализа поступал произносимый речевой фрагмент, а синтез заканчивался озвученным чтением полученного текста на ЕЯ.

Принципиальная новизна создаваемого ЛП состоит в том,

что формальная модель языка, лежащая в его основе, является наиболее полной моделью такого рода. Это модель класса "Смысл ↔ Текст", идеология которой изложена в [Мельчук, 1974]. Такая модель обеспечивает получение связных синтаксических структур для всех предложений обрабатываемых текстов, независимо от степени их сложности, и переработку текстов на естественном языке без смысловых потерь.

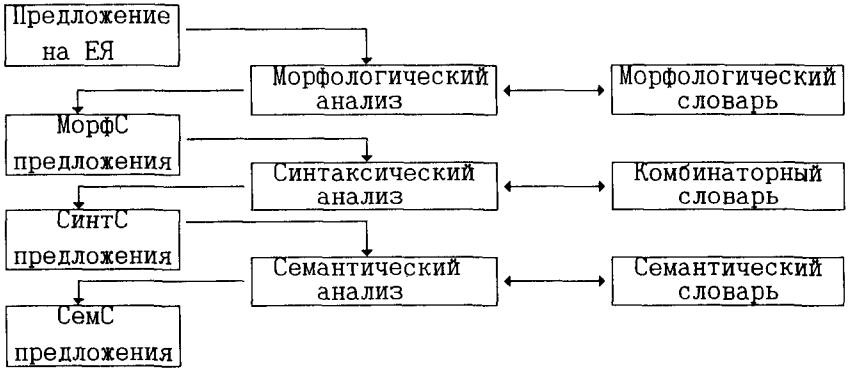
ЛП для общения с БД на естественном языке активно разрабатываются рядом западных фирм (см.: [Kahn, 1984; Sordi, 1984; Lehmann et al., 1985; Эршил, 1988]). Рабочие языки в соответствующих проектах - английский и немецкий. Эти проекты уступают настоящему проекту в силу того, что не опираются на принципиальную лингвистическую модель и тем самым изначально не в состоянии обеспечить полноту и надежность обработки естественно-языковых данных. Кроме того, они не работают с русским языком. Таким образом, данный ЛП является приоритетной разработкой в обоих указанных смыслах.

То же верно и относительно систем машинного перевода. В настоящее время в мире существует рынок коммерческих систем МП (см., напр., [Кулагина, 1989]), которые, однако, не обеспечивают такого высокого качества перевода, как система, построенная на базе рассматриваемого ЛП. Кроме того, в соответствующих разработках практически полностью игнорируется русский язык. Таким образом, и в области МП рассматриваемый ЛП является приоритетной разработкой.

1.2. Структура и состав ЛП

Со стороны своего внутреннего устройства ЛП представляет собой многоуровневый преобразователь. В нем различаются три уровня пофразного представления текста - морфологический, синтаксический и семантический. Каждый из уровней обслуживается соответствующим компонентом модели - массивом правил и определенным словарем или словарями. На каждом из уровней предложение имеет формальный образ, именуемый в дальнейшем его структурой - морфологической (МорфС), синтаксической (СинтС) и семантической (СемС); в примерах, если это не мешает пониманию существа дела, мы позволяем себе опускать некоторые элементы этих структур.

В целом работу модели при анализе можно представить следующей блок-схемой:



Синтез представляет собой обратный переход от СемС предложения к его записи в обычном орфографическом виде. Поскольку процедура анализа, во всяком случае "с точки зрения" компьютера, существенно сложнее процедуры синтеза и заведомо включает в себя все средства, необходимые для этой последней, мы в дальнейшем сосредоточимся преимущественно на анализе.

Под *морфологической структурой понимается последовательность* входящих в анализируемое предложение слов с указанием части речи и морфологических характеристик (падежа, числа, рода, одушевленности, времени, вида и т. п.).

Под *синтаксической структурой понимается* дерево зависимостей, в узлах которого стоят слова данного естественного языка с указанием части речи и грамматических характеристик, а дуги соответствуют специфичным для данного естественного языка отношениям синтаксического подчинения. В описываемой модели используется 40-60 различных отношений; русский синтаксис описывается с помощью 55 отношений.

Под *семантической структурой понимается дерево* зависимостей, в узлах которого стоят либо предметные имена, либо слова универсального семантического языка (например, имена таблиц, в которых сосредоточены сведения о данной предметной области, атрибуты таблиц, операторные символы), а дуги соответствуют универсальным отношениям семантического подчинения, таким, как аргументное, атрибутивное, конъюнкция, дизъюнкция, равенство, неравенство, больше, меньше, принадлежит, не принадлежит и т. п. Существенным компонентом СемС

ляется информация о кореферентности узлов, т.е. информация о том, в каких случаях речь идет об одном и том же объекте, а в каких - о разных. Так, в предложении

2) ***Меня познакомили с человеком, брат которого женат на своей троюродной сестре***

форма ***которого*** обозначает тот же самый объект, что и форма ***человеком,*** а форма ***своей*** - тот же объект, что форма ***брат.*** Кореферентная информация весьма существенна для понимания любых естественно-языковых текстов, включая тексты запросов к базам данных; ср. запросы типа

3) ***Укажите аналитиков отдела сбыта и его менеджера,***

где местоимение ***его*** кореферентно с существительным ***отдел.***

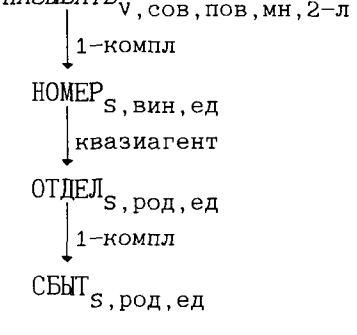
Рассмотрим все три типа структур на примере простого предложения (4):

4) ***Назовите номер отдела сбыта.***

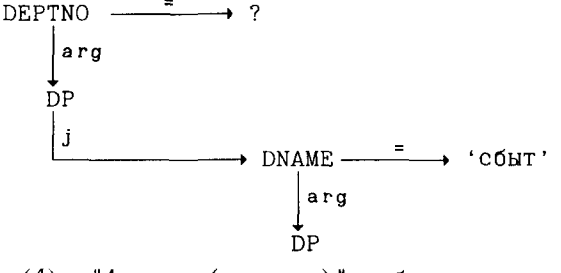
МорфС (4) НАЗЫВАТЬ$_{V,сов,пов,мн,2-л}$ + НОМЕР$_{S,вин,ед}$ + ОТДЕЛ$_{S,род,ед}$ + СБЫТ$_{S,род,ед}$

Для простоты мы здесь не показываем лексико-грамматическую омонимию, например, тот факт, что форма ***номер*** представляет не только винительный, но и именительный падеж данного существительного.

СинтС (4)



СемС (4)



В СинтС (4) "1-компл(етивное)" обозначает отношение,

связывающее слово со вторым актантом (именем участника) обозначаемой им ситуации, а "квазиагент(ивное)" - отношение, связывающее существительное с его первым актантом в форме родительного падежа.

СемС (4) читается следующим образом: "Каков номер отдела, такого, что название этого отдела - 'сбыт'"? Заметим, что ЛП способен поставить в соответствие множеству синонимичных запросов, сформулированных на русском языке, одно и то же выражение на языке семантических структур.

Лингвистический процессор в целом должен обеспечивать выполнение следующих преобразований:

предложение на ЕЯ ⇒ МорфС ⇒ СинтС ⇒ СемС (при анализе),

СемС ⇒ СинтС ⇒ МорфС ⇒ предложение на ЕЯ (при синтезе).

Выше мы уже говорили, что в данной монографии основное внимание будет уделено анализу как более трудной из этих двух процедур. Добавим к этому, что именно процедура анализа обеспечена всеми необходимыми для ее выполнения типами правил. Что касается процедуры синтеза, то лингвистически этап СемС ⇒ СинтС еще не обеспечен.

Итак, чтобы построить ЛП указанного типа, необходимо разработать 1) формальные языки для записи (образов) предложений на морфологическом, синтаксическом и семантическом уровнях представления; 2) формальное понятие структуры предложения для каждого из этих уровней; 3) массивы правил для преобразования структур смежных уровней друг в друга; 4) морфологический, комбинаторный и семантический словари, включив в них всю информацию о каждой лексеме, необходимую для осуществления соответствующего преобразования.

Чтобы получить многоязычный ЛП, такую работу следует выполнить для каждого из участвующих в нем ЕЯ.

Наконец, чтобы получить полифункциональный ЛП, необходимо постоянно пополнять его средствами решения каждой очередной задачи, если они специфичны для нее. Так, для автоматизации перевода с одного ЕЯ на другой ЛП должен быть дополнен соответствующим массивом правил перевода.

1.3. Возможные прикладные функции ЛП

Описываемый ЛП может быть непосредственно использован по крайней мере в двух типах прикладных систем:

1) в системе общения с компьютером на практически не-

ограниченном ЕЯ. Подобная система позволяет организовать максимально дружественный интерфейс пользователя с компьютером, поскольку не требует или почти не требует от пользователя предварительных знаний в области техники и эксплуатации ЭВМ;

2) в системах МП научно-технических текстов и деловой документации с иностранных языков на русский и с русского на иностранные.

В принципе возможно превращение системы общения на ЕЯ в многоязычную путем ее сопряжения с системой МП.

Обе названные задачи ставятся как задачи перевода, с той разницей, что в первом случае речь идет о переводе с ЕЯ на искусственный язык, а именно, на формальный язык запросов, принятый в данной системе управления базами данных (СУБД), а во втором - о переводе с одного ЕЯ на другой.

Имеется еще два класса задач, разрешимых в более отдаленной перспективе, для которых ЛП может оказаться полезным в полном своем объеме.

Первый из этих классов задач связан с автоматическим пополнением баз данных непосредственно по текстам. Для решения таких задач необходимы, помимо собственно лингвистического процессора, способного понимать тексты данной предметной области, еще и логические процессоры. Они должны быть оснащены таким набором функций, с помощью которых можно сравнивать уже имеющуюся в БД информацию со вновь поступающей и извлекать из обрабатываемого текста принципиально новую информацию,. если она в нем содержится.

Второй класс задач относится к недавно появившейся области, получившей название "планирование текста" (text planning; см.: [Hovy, 1988; Iordanskaja-Polguère, 1988; Carcagno-Iordanskaja, 1989]). Если в задачах автоматического пополнения БД по текстам ЛП используется как анализатор, то в задачах планирования текста на первый план выходят его активные - синтезирующие или текстопорождающие - функции. Одна из типичных конкретных задач этого рода - порождение описания на ЕЯ текущей работы СУБД по таким параметрам, как даты и времена ее использования, имена пользователей, которым были оказаны информационные услуги, существо этих услуг (что именно интересовало пользователя), вероятные цели обращения к СУБД и т.п. В этой задаче тоже необходимо сопря-

жение ЛП с логическими процессорами. При этом сначала работают именно логические процессоры, порождающие план текста, а затем - его развернутую концептуальную структуру. После этого вступает в действие ЛП, который сначала преобразует концептуальную структуру в семантическую структуру будущего текста на ЕЯ, а затем, через ряд промежуточных этапов, превращает ее в последовательность предложений, образующих связный рассказ на заданную пользователем тему.

В обоих указанных классах задач предполагается использование ЛП в полном объеме. Этим его возможные функции и применения не исчерпываются. Различные элементы и модули лингвистического процессора, прежде всего, содержащиеся в нем лингвистические знания, допускают использование и в других информационных системах, прежде всего, в партнерских системах типа "помощник учителя" (русского или иностранных языков), а также в разного рода компьютерных словарях.

Рассмотрим более подробно две задачи, сформулированные в начале настоящего раздела, для принципиального решения которых достаточно средств самого ЛП в его нынешнем виде, - задачу общения с базами данных на естественном языке и задачу машинного перевода.

В зависимости от характера конкретной решаемой задачи переработка предложения доводится до большего или меньшего уровня глубины. Так, в задаче общения с БД на неограниченном ЕЯ необходимо считаться с тем, что формальный образ запроса на языке СУБД может разительно отличаться от исходного предложения на ЕЯ. Например, запрос

(5) ***Сколько клерков работает в коммерческом отделе, и какова их суммарная зарплата?***

"в переводе" на SQL - один из широко применяемых языков запросов к реляционным базам данных - приобретет следующий вид:

```
(5')   SELECT   COUNT (ENAME), SUM (SAL)
       FROM     EM
       WHERE    DEPTNO = ( SELECT    DEPTNO
                           FROM      DP
                           WHERE (DNAME = 'коммерческий'))
                AND   JOB = 'клерк'
```

Очевидно, что формальный образ запроса очень далеко отстоит от своего прототипа на ЕЯ. Поэтому для осуществления

перевода (5) в (5') необходимо произвести очень глубокую переработку предложения - достичь такого уровня представления смысла, который является универсальным для всех естественных языков и любых формальных, если данный смысл в принципе может быть в них выражен. Практически это значит, что должен быть достигнут уровень СемС.

В задаче перевода с одного ЕЯ на другой столь глубокая переработка текста не нужна. Действительно, предложение (5) в переводе на английский язык будет выглядеть так:

(5'') How many clerks are working in the commercial department and what is their total salary?

Совершенно очевидно, что это предложение по своим морфологической и синтаксической структурам, порядку слов и т.п. гораздо ближе к (5), чем предложение на SQL. Поэтому при переводе с одного ЕЯ на другой нет необходимости достигать максимального уровня глубины анализа. Можно воспользоваться сходством синтаксических структур переводящих друг друга предложений естественных языков и произвести изменения лишь в таких точках исходного предложения, где наблюдается морфологическая, синтаксическая или лексическая национальная специфика. Снятие такой специфики обеспечивается специальным массивом правил перевода, на вход которых поступает СинтС русского (или английского) предложения, а на выходе получается СинтС английского (или русского) предложения. Сами правила сначала снимают национальную специфику входного языка относительно выходного, а затем порождают национальную специфику выходного языка.

1.4. Принципы разработки ЛП

При разработке ЛП для сложных информационных систем мы руководствовались следующими принципами.

1. Независимость форматов задания лингвистических знаний от языка, которая обеспечивает возможность их наполнения данными любых языков и превращения любых систем в многоязычные.

2. Независимость грамматик и словарей от алгоритмов, которая обеспечивает "открытость" лингвистических знаний и возможность удобной корректировки грамматик и словарей. По существу этот принцип означает "декларативность" задания лингвистической информации. Он последовательно проведен во

всей системе, за исключением части алгоритма синтаксического анализа предложения: наряду с полным алгоритмом синтаксического анализа для ускорения работы используется алгоритм фрагментного анализа (см. о нем разд. 4.8.2), в рамках которого существенная часть необходимых для него лингвистических знаний задается процедурно.

3. Независимость лингвистических знаний от предметной области, благодаря которой обеспечивается возможность адаптации единых алгоритмов анализа и синтеза текстов для обработки данных из новых предметных областей.

4. Независимость лингвистических знаний от характера решаемой задачи, которая позволяет использовать одну и ту же формальную модель языка для решения самых различных задач, например, таких, как общение с БД на ЕЯ, с одной стороны, и МП - с другой. Лингвистический процессор, удовлетворяющий этому требованию, оказывается способным обслуживать многие типы сложных информационных систем, т.е. приобретает черты полифункциональности.

Выполнение двух последних условий означает, что формальная модель языка, лежащая в основе ЛП, должна быть теоретически обоснованной (построенной с учетом всех новейших достижений лингвистической науки) и полной. ЛП должны быть доступны знания ЕЯ в объеме, в котором им владеют его носители. Практически это значит, что ЛП должен быть способен обрабатывать любые морфологические явления данного ЕЯ, все синтаксические конструкции деловой прозы и около 10 000 слов общего распространения, которые встречаются в любых типах текстов на данном языке. При этом ЛП должен одинаково хорошо владеть как внешней стороной всех перечисленных языковых средств, так и их семантикой. Только такая принципиальная модель языка может обеспечить переработку произвольных текстов, не требуя коренной перестройки при переходе к каждому новому типу текстов.

Современное состояние лингвистических знаний позволяет строить стопроцентно полные и надежные модели морфологии. Достаточной степени полноты и надежности можно добиться и в модели синтаксиса. Совершенно реальной перспективой является создание словарей объемом до нескольких десятков тысяч слов, обеспечивающих высококачественный морфологический и синтаксический анализ и синтез текстов деловой прозы на

практически неограниченном ЕЯ. Это открывает возможность сделать морфологический и синтаксический компоненты процессора универсальными, т. е. не зависящими от характера конкретной прикладной задачи. Они могут использоваться в широком спектре информационных систем, включая системы автоматического понимания текстов, МП, общения с БД и т. п., т. е. действительно приобретают свойства полифункциональности. Если же говорить о семантике, то в своем нынешнем состоянии она еще не готова к построению универсальных моделей своего объекта, т. е. моделей, учитывающих настолько широкий круг явлений ЕЯ, что они способны обслуживать любые прикладные задачи. Поэтому модель семантики в нашем ЛП была с самого начала ориентирована на конкретную проблему. О том, как это было сделано, см. гл. 5.

1.5. Источники разработки и компоненты ЛП

Отправным пунктом для работы над ЛП в том виде, как он описан выше, послужили более конкретные и предметно-ориентированные исследования в области МП, проводившиеся нашей лингво-математической группой с конца 70-х до середины 80-х годов. Они описаны с достаточной степенью подробности в монографии [Апресян и др., 1989]. С 1986 г. на основе этих исследований развернулась работа над ЛП. Однако в качестве экспериментального полигона было решено использовать не только систему общения с БД на ЕЯ, но и системы МП. С этой целью на новой технической основе и со значительно расширенным и усовершенствованным лингвистическим обеспечением была возрождена система англо-русского МП, подробно описанная в упомянутой монографии. Кроме того, были проведены работы по использованию формальной модели русского языка для системы обратного МП с русского языка на английский, завершившиеся серией успешно проведенных экспериментов. Результатом этих работ стала система двунаправленного перевода ЭТАП-3. Наконец, была сделана попытка подсоединить систему МП с английского языка на русский к системе общения с БД на ЕЯ, давшая обнадеживающие результаты. Таким образом, ЛП, задуманный как фундаментальная разработка в области моделирования понимания и производства текстов на ЕЯ, в настоящее время проходит успешную экспериментальную проверку в трех типах сложных информационных систем.

К настоящему времени разработаны все основные элементы лингвистического и логико-алгоритмического обеспечения ЛП.

В области лингвистического обеспечения:

а) разработана форма представления МорфС, СинтС и СемС;

б) построены модели морфологического анализа и синтеза (для русского и английского языков);

в) построены модели синтаксического анализа и синтеза (для русского и английского языков), каждая из которых насчитывает до 500 - 600 правил; для ускорения работы алгоритмов синтаксического анализа и синтеза все правила разделены на общие, включающиеся в обработку любого предложения, и словарные, активируемые определенными словами, если они встретились в предложении; словарные правила делятся на трафаретные, применимые к замкнутому классу слов, и собственно словарные, применимые только к данному слову и записываемые в его словарной статье;

г) построена модель семантического анализа (для русского языка), насчитывающая около 100 правил;

д) построены массивы правил перевода с английского языка на русский и с русского языка на английский, насчитывающие до 500 общих и трафаретных правил каждый, и большое число собственно словарных правил перевода;

е) построены морфологические словари русского и английского языков; первый из них насчитывает до 12 000 слов, а второй - до 15 000; второй словарь используется в подсистеме быстрого машинного перевода, которая выдает пословный перевод обрабатываемого предложения на русский язык вместе с его морфологической структурой;

ж) построены английский и русский комбинаторные словари, предназначенные для систем англо-русского и русско-английского автоматического перевода, объемом около 10 000 единиц каждый;

з) построен семантический словарь для русского языка, предназначенный для перевода запросов на язык SQL.

В области логико-алгоритмического обеспечения:

а) разработаны формальные языки для записи лингвистической информации; все лингвистические знания, используемые в ЛП, записаны на этих формальных языках;

б) построены алгоритмы морфологического анализа и синтеза, обеспечивающие преобразование предложений на ЕЯ в их

МорфС и обратно;

в) построены алгоритмы предсинтаксического и синтаксического анализа текстов, получающие на входе МорфС обрабатываемого предложения и выдающие на выходе его СинтС;

г) построен алгоритм преобразования древесных структур, обеспечивающий последовательное применение к структуре всех необходимых лингвистических правил ее преобразования для получения СинтС выходного языка (в системах МП) или СемС (в системе общения с БД на ЕЯ);

д) построен алгоритм перевода СемС в формулу языка SQL.

Разработаны комплексы программ, обеспечивающие трансляцию лингвистических данных в машинную форму и реализующие все упомянутые выше алгоритмы.

Все работы проводились на ЭВМ VAX-75Ø; программы написаны на языке PL/1; время обработки одного предложения - от 5 до 20 сек для перевода с русского языка на SQL и от 20 до 80 сек для перевода с английского языка на русский или с русского на английский.

К 1990 г. были построены экспериментальная система общения с БД на практически неограниченном русском языке и экспериментальная система МП научно-технических текстов с русского языка на английский, а также возрождена и усовершенствована созданная ранее экспериментальная система англо-русского МП.

В последующих главах принят следующий порядок изложения. В гл. 2 описываются формальные языки для записи лингвистической информации. Главы 3-5 посвящены лингвистическим компонентам и алгоритмическим модулям ЛП. Они излагаются в том порядке, в котором включаются в работу алгоритмом. В гл. 6 описываются два основных словаря ЛП - комбинаторный и семантический. Последняя, седьмая глава посвящена описанию экспериментов с ЛП в системах общения с базами данных на ЕЯ и системах машинного перевода. В Заключении намечаются некоторые перспективы дальнейшей работы над лингвистическим процессором.

С целью сделать данную работу максимально независимой от других наших публикаций на связанные с ЛП темы мы пошли на некоторое дублирование информации. Впрочем, почти нигде она не воспроизводится дословно, поскольку ЛП в процессе работы над ним постоянно развивался и совершенствовался.

Глава 2
ФОРМАЛЬНЫЕ ЯЗЫКИ ДЛЯ ЗАПИСИ ЛИНГВИСТИЧЕСКОЙ ИНФОРМАЦИИ

Формальные языки, используемые для записи лингвистической информации в ЛП, уже были предметом описания (см.: [Цинман, 1986 а, б; Райхлина, Цинман, 1986; Апресян и др., 1989]). Однако для целей настоящей работы эти языки были серьезно расширены и усовершенствованы с учетом современных потребностей ЛП, в особенности потребностей, связанных с включением в процесс переработки текстов на ЕЯ семантических правил. Появились, в частности, десять новых предикатов и десять новых инструкций.

Стремясь обеспечить данной книге максимальную автономность, мы опишем используемые в ЛП формальные языки не выборочно (только новые предикаты, только новые инструкции), а целиком - в том виде, в каком они используются на нынешнем этапе существования системы.

В ЛП имеются формальные языки для записи трех основных видов лингвистической информации: 1) статей морфологического словаря; 2) статей комбинаторного словаря; 3) синтаксических и трансформационных правил.

Статьи морфологического словаря имеют стандартную структуру, позволяющую описывать в единой и компактной форме морфологию не только русского, но и других потенциальных рабочих языков ЛП.

Статьи комбинаторного словаря тоже записываются на едином формальном языке, рассчитанном на материалы любых ЕЯ. Статьи состоят из зон двух типов: классификационных и операционных.

Запись классификационных зон статьи комбинаторного словаря не представляет особых формальных сложностей, и поэтому их изложение дается там же, где описывается соответствующая лингвистическая информация (см. гл. 6).

Операционные зоны заполняются либо конкретными словарными правилами, либо ссылками на трафаретные правила.

Правила - наиболее сложные формальные объекты в нашей системе. Практически вся нетривиальная лингвистическая ин-

формация записана в виде разного рода правил, используемых на всех этапах работы, кроме морфологических.

Создание подходящего формального языка для записи этой информации явилось одной из наиболее принципиальных проблем при построении ЛП. Дело в том, что этот язык должен быть в равной степени удобным как для лингвистов, так и для программистов.

Лингвистам нужен формальный язык, который обладал бы богатым набором выразительных средств (поскольку на нем приходится записывать весьма тонко организованную информацию), был бы достаточно свободным (нельзя заранее предугадать, какого рода лингвистические сведения могут в дальнейшем понадобиться) и близким к естественному (важное требование, так как при создании любой системы ЛП неизбежен длительный период отладки лингвистического обеспечения, которая осуществляется самими лингвистами в терминах формального языка). Основное требование программистов к формальному языку - максимально возможная простота алгоритмической работы с выражениями этого языка. Эти требования весьма противоречивы, и от того, найдется ли между ними разумный компромисс, зависит успех всей системы в целом.

Нам представляется, что предлагаемый формальный язык удовлетворяет перечисленным требованиям. Это в значительной степени объясняется тем обстоятельством, что он создавался с учетом всего накопленного опыта и на основе детального анализа имевшихся к тому моменту содержательных описаний грамматик и комбинаторных словарей русского, английского и французского языков, выполненных в рамках модели "Смысл ↔ Текст".

2.1. Структура правила

Каждое правило начинается с заголовка, назначение которого - идентификация правила (отсылка к тому или иному правилу в системе производится по его заголовку). Заголовок имеет вид

REG ИМЯ N,

где ИМЯ - имя совокупности правил, которой принадлежит данное правило, а N - номер данного правила в этой совокупности. На каждом этапе работы системы происходит обращение к соответствующим совокупностям правил.

Правила в системе ЛП бывают двух видов: элементарные и обобщенные (альтернативные).

Элементарное правило состоит из двух зон: зоны CHECK (проверить), содержащей список условий, и зоны DO (исполнить), содержащей список действий (инструкций). Если условия из зоны CHECK оказываются выполненными, то правило считается применимым и совершаются действия, перечисленные в зоне DO.

Обобщенное правило состоит из зоны общих условий и нескольких элементарных подпра́вил. Работа с обобщенным правилом проводится следующим образом. Сначала проверяются общие условия правила. К подправилам обращаются лишь в том случае, если выполнены все общие условия. Предполагается, что подправила обобщенного правила являются альтернативными. Это значит, что выполнение условий одного из подправил освобождает нас от необходимости работать с остальными и предполагает переход к зоне DO данного подправила. Многие элементарные правила естественным образом "склеиваются" в обобщенные; это заметно уменьшает как общее число правил в системе, так и время просмотра всей совокупности правил. Кроме того, обобщенные правила оказались удобным средством записи некоторых сложных логических выражений, которые иным способом на нашем формальном языке записать не удается.

Помимо указанных двух зон, в правиле может записываться информация об использованных в нем термовых переменных и их значениях, а также особые указания алгоритму.

2.2. Сигнатура формального языка

В нашем языке все условия в зоне CHECK представляют собой выражения логики предикатов первого порядка над некоторой сигнатурой. Действия из зоны DO записываются в виде операторов (инструкций) над этой сигнатурой.

2.2.1. Термы

Все термы в сигнатуре языка заданы перечнем списков. Каждый терм в машинном виде представлен двумя числами: номером списка и порядковым номером в этом списке.

2.2.1.1. Предметные константы

В сигнатуре представлено около 20 списков предметных констант. Это морфологические характеристики (отдельные

списки для характеристик части речи, рода, числа, падежа, лица, времени, наклонения, вида, залога, степеней сравнения), синтаксические признаки, семантические признаки (дескрипторы), знаки препинания, имена синтаксических отношений, имена лексем, статьи которых представлены в комбинаторных словарях, и т. д. Каждый из этих списков в отдельности и множество списков в целом остаются открытыми для пополнения и изменения.

Сигнатура содержит около 600 термов, не являющихся лексемами, и около 20 000 имен лексем (английских и русских). Распределение термов по спискам не случайно. В предикатах и инструкциях широко используется принадлежность термов одному списку (см., например, предикаты согласования и инструкции, работающие с характеристиками слов). Кроме того, многие списки содержат термы, обобщенные для данного списка. Эти обобщенные термы удобны в том случае, когда пользователь не хочет или не может уточнить, о каком конкретном терме данного списка идет речь. С помощью обобщенных термов можно, например, перенести от одного слова к другому характеристику числа (падежа, рода и т. д.) без выяснения, какова эта характеристика.

2.2.1.2. Предметные переменные

В языке имеются два вида предметных переменных: контекстные (узловые) и термовые (обобщающие). Значениями контекстных переменных являются слова обрабатываемой фразы (точнее, номера слов фразы с указанием номера омонима, если слово имело омонимичные разборы). Переменная X выделяет некоторое слово, а другие переменные (обозначаемые через Y, Z, U, W, Q, возможно, с цифровыми индексами) используются для описания контекста этого слова.

Термовые переменные предназначены для обобщения правил и, по существу, превращают их в схемы правил. Значениями этих переменных являются термовые константы. При использовании термовой переменной список ее значений помещается в тексте правила. В языке для обозначения термовых переменных используются следующие имена: ALFA, BETA,..., R, R1, R2,..., - для неповторимых переменных; RALFA, RBETA,..., RR, RR1, RR2,... - для повторимых термовых переменных. Информация о повторимости (т. е. о наличии нескольких вхожде-

ний одной переменной в правило) используется алгоритмической процедурой работы с правилом.

2.2.2. Предикаты

В сигнатуре языка имеется 54 элементарных и около 200 составных предикатов. Все предикаты записываются в удобной для транслятора префиксной форме. Предполагается, что для всех предикатов указаны число мест ("арность") и область определения каждого аргумента, которая задается перечислением номеров списков термов. Эта информация используется транслятором правил для контроля правильности заполнения предикатных мест. Совокупность предикатов может быть расширена. Для этого, помимо введения в сигнатуру нового предикатного имени, требуется создать и зарегистрировать в системе алгоритмическую процедуру, вычисляющую истинность данного предиката.

2.2.2.1. Элементарные предикаты

Список элементарных предикатов условно можно разбить на пять групп: 1) предикаты идентификации, 2) предикаты линейного порядка, 3) предикаты доминации, 4) предикаты согласования, 5) предикаты моделей управления. В записи предикатов используются следующие обозначения: Z, Z1, Z2,... - контекстные переменные; t_1, t_2,...,l_1, l_2 ,..., n, r – предметные константы или термовые переменные. Говоря о значениях контекстных переменных, будем их иногда называть словоформами, или словами (в предикатах идентификации, согласования и линейного порядка), узлами (в предикатах доминации), лексемами (в предикатах, работающих с моделями управления). Это связано с желанием подчеркнуть существенную для рассматриваемого предиката роль значения переменной: иметь набор характеристик, занимать определенное место во фразе, быть узлом в структуре, иметь модель управления и т. д.

Предикаты идентификации (1-9)

Все предикаты этой группы, а также некоторые предикаты в других группах являются переменноместными.

(1) $=(Z,t_1,\ldots,t_k)$ или $EQU(Z,t_1,\ldots,t_k)$, где $k>1$, - две возможные записи предиката "словоформа Z обладает всеми характеристиками $t_1,\ldots,t_k$".

(2) $\#(Z,t_1,\ldots,t_k)$ или $NEQUN(Z,t_1,\ldots,t_k)$, где $k>1$, -

две возможные записи предиката "словоформа Z не обладает ни одной из характеристик $t_1,\ldots,t_k$".

(3) EQUN(Z,$t_1,\ldots,t_k$), где k>1, - "словоформа Z обладает хотя бы одной из характеристик $t_1,\ldots,t_k$".

(4) NEQU(Z,$t_1,\ldots,t_k$), где k>1, - "словоформа Z не обладает хотя бы одной из характеристик $t_1,\ldots,t_k$".

(5), (6) LEXA(Z,$l_1,\ldots,l_k$) <LEXR(Z,$l_1,\ldots,l_k$)>, где k>1, - "Z - словоформа, принадлежащая одной из английских <русских> лексем $l_1,\ldots,l_k$".

(7), (8) NLEXA(Z,$l_1,\ldots,l_k$) <NLEXR(Z,$l_1,\ldots,l_k$)>, где k>1, - "Z не является словоформой ни одной из английских <русских> лексем $l_1,\ldots,l_k$".

(9) HOM(Z,$t_1,\ldots,t_k$), где k>Ø, - "словоформа Z омонимична, и хотя бы один из ее омонимов обладает всеми характеристиками $t_1,\ldots,t_k$"; при k=Ø HOM(Z) истинен, если у Z есть хотя бы один омоним.

Предикаты линейного порядка (10-28)

(10) ORD(Z,Z1,Z2,...) - "слово Z во фразе расположено левее Z1, Z1 левее Z2 и т.д.".

(11) NEXT(Z,Z1,Z2,...) - "слово Z - непосредственный левый сосед Z1, Z1 - непосредственный левый сосед Z2 и т.д.".

(12) L(Z,Z1,n) или DIST(Z,Z1,n) - две возможные записи предиката "слово Z расположено левее Z1, и между ними находится не более чем n слов".

(13) R(Z,Z1,n) - "слово Z расположено правее Z1, и между ними находится не более чем n слов".

(14) M(Z,Z1,n) - "слово Z отстоит от Z1 не более чем на n слов (влево или вправо)".

(15) I(Z,Z1,Z2) - "слово Z1 находится между Z и Z2".

(16), (17) LEFT(Z,$t_1,\ldots,t_k$) <RIGHT(Z,$t_1,\ldots,t_k$)>, где k>Ø, - "слева <справа> от Z есть словоформа, обладающая всеми характеристиками $t_1,\ldots,t_k$"; при k = Ø LEFT(Z) <RIGHT(Z)> истинен, когда Z не является первым <последним> словом фразы.

(18), (19) NLEFT(Z,$t_1,\ldots,t_k$) <NRIGHT(Z,$t_1,\ldots,t_k$)>, где k>Ø, - "непосредственно слева <справа> от Z стоит словоформа, обладающая всеми характеристиками $t_1,\ldots,t_k$".

(20) INSENT($t_1,\ldots,t_k$), где k>Ø, - "во фразе есть слово-

форма, обладающая всеми характеристиками t_1, ,t_k"

(21) ININT(Z,Z1,t_1, ,t_k), где k>Ø, - "в интервале между Z и Z1 есть словоформа, обладающая всеми характеристиками t_1, ,t_k", при k=Ø ININT(Z,Z1) истинен, когда Z и Z1 не соседние слова.

(22) EQUINT(Z,Z1,t_1, ,t_k), где k>1, - "в интервале между Z и Z1 находится ровно k словоформ, первая (слева направо) из которых имеет характеристику t_1, вторая - характеристику t_2, , а последняя - характеристику t_k"

(23), (24) PLEFT(Z,t_1, ,t_k) <PRIGHT(Z,t_1, ,t_k)>, где k>0, - "слева <справа> от Z есть один из знаков препинания t_1, ,t_k", при k=Ø PLEFT(Z) <PRIGHT(Z)> истинен, когда слева <справа> от Z есть хотя бы один знак препинания.

(25), (26) PNLEFT(Z,t_1, ,t_k) <PNRIGHT(Z,t_1, ,t_k)>, где k>0, - "непосредственно слева <справа> от Z есть один из знаков препинания t_1, ,t_k", при k=Ø PNLEFT(Z) <PNRIGHT(Z)> истинен, когда непосредственно слева <справа> от Z есть хотя бы один знак препинания.

(27) PINSENT(t_1, ,t_k), где k>Ø, - "во фразе есть один из знаков препинания t_1, ,t_k".

(28) PININT(Z,Z1,t_1, ,t_k), где k>Ø, - "в интервале между Z и Z1 есть один из знаков препинания t_1, ,t_k"; при k=Ø PININT(Z,Z1) истинен, если в интервале между Z и Z1 есть хотя бы один знак препинания.

Предикаты доминации (29-37)

(29), (30) DOM(Z,Z1,r) <DEP(Z,Z1,r)> - "узел Z является непосредственным синтаксическим хозяином <слугой> узла Z1 по отношению r"

(31), (32) IDOM(Z,Z1,n) <IDEP(Z,Z1,n)> - "узел Z является опосредованным синтаксическим хозяином <слугой> узла Z1, и путь между ними в синтаксической структуре проходит не более чем через n узлов".

(33) HOMDOM1(Z,Z1,r) - "у узла Z есть омоним, являющийся синтаксическим хозяином узла Z1 по отношению r".

(34) HOMDOM2(Z,Z1,r) - "узел Z является синтаксическим хозяином некоторого омонима узла Z1 по отношению r".

(35) HOMDOM12(Z,Z1,r) - "у узла Z есть омоним, являющийся синтаксическим хозяином некоторого омонима узла Z1 по отношению r".

(36) CHAIN(Z1,Z2,r_1, ,r_k) - "от узла Z1 до узла Z2 можно спуститься по дереву, при этом все ветви данного пути отмечены именами синтаксических отношений из списка r_1, ,r_k"

(37) CUMBER(Z,n) - "количество узлов в синтаксической группе узла Z не превосходит n". [1]

Предикаты согласования (38-47)

(38), (39) COCAS(Z1,Z2) <CONMB(Z1,Z2)> - "обе словоформы Z1 и Z2 имеют характеристики падежа <числа>, и эти характеристики совпадают".

(40), (41), (42) COPER(Z1,Z2) <COGEN(Z1,Z2),COAN(Z1,Z2)> - "либо обе словоформы Z1 и Z2 имеют характеристики лица <рода, одушевленности>, и эти характеристики совпадают, либо хотя бы одна из этих словоформ не имеет характеристики лица <рода, одушевленности>".

(43) CODES(Z1,Z2) - "лексемы, которым принадлежат словоформы Z1 и Z2, имеют хотя бы один общий дескриптор, либо хотя бы одна из этих лексем не имеет ни одного дескриптора".

(44) CODESSEM(Z1,Z2) - отличается от предиката CODES (Z1,Z2) лишь тем, что рассматриваются дескрипторы из специального списка, ориентированного на конкретную базу данных.

(45) COLEX(Z1,Z2) - "Z1 и Z2 являются словоформами одной и той же лексемы".

(46) HOMCOCAS(Z1,Z2) - "хотя бы один из омонимов словоформы Z1 имеет ту же характеристику падежа, что и словоформа Z2"

(47) COREF(Z1,Z2) - "Z1 и Z2 кореферентны".

Предикаты моделей управления (48-54)

(48) VAL(n,Z,t_1, ,t_k), где k>1, - "в n-м столбце модели управления лексемы Z есть запись, содержащая все характеристики t_1, ,t_k"

(49) CORDES(n,Z,Z1) - "в n-м столбце модели управления лексемы Z есть запись, содержащая хотя бы один дескриптор, которым обладает лексема Z1, либо в n-м столбце модели

---

[1] Под синтаксической группой узла Z понимается поддерево синтаксической структуры с вершиной Z

управления лексемы Z или у лексемы Z1 дескрипторов нет".

(50) CORSYNT(n,Z,Z1) - "в n-м столбце модели управления лексемы Z есть запись, содержащая синтаксические признаки, которыми обладает лексема Z1".

(51) CORCAS(n,Z,Z1) - "в n-м столбце модели управления лексемы Z есть запись, содержащая характеристики падежа, которыми обладает лексема Z1".

(52) CORLEX(n,Z,Z1) - "в n-м столбце модели управления лексемы Z есть запись, содержащая имя лексемы Z1".

(53) HOMCORCAS1(n,Z1,Z2) - "хотя бы один из омонимов словоформы Z1 имеет в n-м столбце модели управления запись, содержащую характеристику падежа, совпадающую с характеристикой падежа словоформы Z2".

(54) HOMCORCAS2(n,Z1,Z2) - "у словоформы Z1 в n-м столбце модели управления есть запись, содержащая характеристику падежа, совпадающую с характеристикой падежа одного из омонимов словоформы Z2".

2.2.2.2. Составные предикаты

Рассмотрим стандартный пример введения новой переменной при описании контекста словоформы X: "справа от X на расстоянии не более n слов найдется словоформа Z, являющаяся синтаксическим хозяином X по отношению r и обладающая одной из характеристик $t_1, t_2, \ldots, t_k$" Это утверждение может быть записано в виде конъюнкции трех элементарных предикатов из числа перечисленных выше. В нашем формальном языке имеется возможность записать это выражение одним составным предикатом, имя которого набирается из имен предикатов, входящих в конъюнкцию: R-DEP-EQUN(X,Z,n,r,$t_1, t_2, \ldots, t_k$).

Составной предикат истинен, если найдется словоформа Z со всеми перечисленными свойствами.

Аппарат составных предикатов, с одной стороны, естествен и удобен, а с другой - заметно ускоряет процедуру поиска требуемого значения вводимой переменной, поскольку в рамках одного составного предиката поиск этого значения можно оптимизировать, начав его с проверки более редких свойств.

Для получения составных предикатов разрешается использовать следующие три списка элементарных предикатов:

1) предикаты линейного порядка:

L(Z,Z1,n); R(Z,Z1,n); M(Z,Z1,n); I(Z,Z1,Z2);

2) предикаты доминации:

DOM(Z,Z1,r); DEP(Z,Z1,r); IDOM(Z,Z1,n); IDEP(Z,Z1,n);

3) предикаты идентификации:

EQU(Z1,t_1,...,t_k); EQUN(Z1,t_1,...,t_k); NEQU(Z1,t_1,...,t_k); NEQUN(Z1,t_1,...,t_k); LEXA(Z1,l_1,...,l_k); LEXR(Z1,l_1,..., l_k); NLEXA(Z1,l_1,..., l_k); NLEXR(Z1,l_1,...,l_k).

Имя составного предиката набирается из имен элементарных предикатов этих списков (по одному из каждого списка). Имена элементарных предикатов записываются в составном имени в том же порядке, как и представленные здесь списки.

Примечания.

1. Имя составного предиката может составляться не из трех, а из двух имен элементарных предикатов.

2. Если для описания линейного порядка использован предикат I(Z,Z1,Z2) ("Z1 между Z и Z2"), то возможны два случая: слово Z1 синтаксически связано либо с Z, либо с Z2. Для различения этих случаев применяются разные записи составного предиката: I-DOM1(Z,Z1,Z2,r) и I-DOM2(Z,Z1,Z2,r).

2.2.3. Инструкции и параметры

Помимо термов и предикатов, сигнатура языка содержит еще два вида объектов: инструкции и параметры. Они рассматриваются в разд. 2.3.

2.2.4. Анонимность неповторимых переменных

Многие контекстные переменные встречаются в записи правил лишь один раз. Такие переменные будем называть неповторимыми. Этим переменным можно не присваивать имена, а на соответствующих местах в предикатах ставить *. Символом * можно заменять в предикатах и некоторые предметные константы. Например, запись вида R-DEP-EQUN(Z,*,*,*,t_1,...,t_k) означает "справа от слова Z найдется некоторая словоформа, отстоящая от него на неопределенном расстоянии, являющаяся его синтаксическим хозяином по некоторому отношению и обладающая одной из характеристик t_1,...,t_k".

2.3. Запись условий в зоне проверки

2.3.1. Дизъюнктивная нормальная форма (ДНФ)

Каждое условие в правилах есть логическое выражение над описанной сигнатурой, представленное в ДНФ (т.е. имеет вид

дизъюнкции конъюнкций предикатов или их отрицаний). Это ограничение, безусловно, упрощает алгоритмическую процедуру проверки истинности таких выражений. С другой стороны, оно оказалось не обременительным для пользователей языка (в противном случае процесс получения ДНФ можно было бы автоматизировать).

2.3.2. Необходимые и невозможные условия

Совокупность условий всякого правила описывает контекст некоторой словоформы X. Это описание может состоять из двух частей: 1) описание того, что должен содержать этот контекст, и 2) указание того, чего в нем быть не может. Соответственно все условия правила бывают двух видов: необходимые и невозможные.

2.3.3. Группы условий

Все условия в правилах делятся на группы. Каждое условие имеет два номера: номер группы, которой оно принадлежит, и порядковый номер в этой группе. В правилах допускается 4 группы условий. Группы с номерами 1, 3 содержат необходимые условия, а с номерами 2, 4 - невозможные условия. Распределение условий на 4, а не на 2 группы позволяет передать алгоритмической процедуре дополнительную информацию (см. описание алгоритма синтаксического анализа в разд. 4.8.1).

2.3.4. Выделенная переменная X

В нашей системе условия во всех правилах описывают контекст некоторого слова, обозначаемого через X. Значение этой переменной подбирается процедурой, внешней по отношению к правилу, и тем самым при работе с правилом не варьируется. Для общих правил в качестве значения X перебираются все словоформы фразы, для словарных и трафаретных правил в качестве значения X рассматривается только ключевое слово словарной статьи, где встретилось это правило или ссылка на него.

2.3.5. Выделенная переменная Y

Алгоритм предусматривает также особую работу с переменной Y. Поясним эту особенность. Пусть условия некоторого правила, описывающие контекст словоформы X, оказались истинными. Тогда, как это предусмотрено, выполняются дейст-

вия, указанные в зоне DO. Однако на этом работа с правилом не заканчивается. Просматривается перечень встретившихся в правиле переменных. Если среди них есть Y, то осуществляется возврат к зоне CHECK с целью проверить истинность условий с каким-либо иным значением Y (а значит, и других переменных, зависящих от Y). Если это удалось, то опять происходит обращение к зоне DO, и т. д.

Необходимость в использовании переменной Y возникает прежде всего в синтагмах, где словоформу X надлежит соединить синтаксической связью со всеми такими словоформами Y, для которых выполняются условия контекста.

2.4. Запись инструкций в зоне действий

Рассмотрим еще один список формальных объектов, входящих в сигнатуру языка: список инструкций, назначение которых - описать требуемые правилом преобразования За каждой инструкцией стоит алгоритмическая процедура, выполняющая действия, предусмотренные этой инструкцией. Для всякой инструкции, как и для предиката, задается число мест ("арность") и область определения каждого аргумента Зона DO в правилах представляет собой перечень инструкций, который фиксирует последовательность требуемых действий. Инструкции могут содержать переменные обоих типов, введенные в зоне CHECK Некоторые инструкции могут вводить новую контекстную переменную, которая затем может повторяться в последующих инструкциях этого правила. Таким способом, например, к фразе добавляется новое слово, которое потом вводится в синтаксическую структуру и наделяется нужными характеристиками.

В настоящий момент в ЛП действуют 45 инструкций, которые разбиты на семь групп в соответствии с выполняемыми ими функциями:

1) работа с характеристиками слов (добавление, стирание, изменение, перенесение);

2) изменение синтаксической структуры фразы;

3) изменение линейного порядка слов и синтаксических групп во фразе;

4) изменение словарного состава фразы (удаление и добавление слова, замена одного слова другим),

5) обработка знаков препинания;

6) работа с кореферентностью;

7) вспомогательная инструкция.

Совокупность инструкций может пополняться так же, как и список предикатов. Для введения новой инструкции надо написать соответствующую алгоритмическую процедуру и зарегистрировать ее в системе.

В описании инструкций символы Z, Z1,... обозначают контекстные переменные; символы $t_1,\ldots, t_k$, l, r, r_1 – предметные константы или термовые переменные.

Инструкции для работы с характеристиками слов (1-4)

(1) DOBUZHAR:Z($t_1,\ldots,t_k$) – "добавить узлу Z характеристики $t_1,\ldots,t_k$".

(2) STERUZHAR:Z($t_1,\ldots,t_k$) – "стереть у узла Z характеристики $t_1,\ldots,t_k$".

(3) ZAMUZHAR:Z($t_1,\ldots,t_k$) – "заменить у узла Z имеющиеся характеристики характеристиками $t_1,\ldots,t_k$ из соответствующих списков термов".

(4) PERUZHAR:Z($t_1,\ldots,t_k$) – "перенести все характеристики из списков, которым принадлежат термы $t_1,\ldots,t_k$, от узла Z к узлу Z1".

Инструкции, изменяющие синтаксическую структуру фразы (5-12)

(5) SVUZOT:(Z,Z1,r) – "связать Z (хозяина) и Z1 (слугу) отношением r".

(6) SVUZOTOK:(Z,Z1,r) – "связать Z (хозяина) и Z1 (слугу) отношением r; связь объявить окончательной".

(7) STEROT:(Z,Z1,r) – "стереть отношение r, связывающее Z (хозяина) и Z1 (слугу)".

(8) IZOT:(Z,Z1,r_1) – "изменить имя отношения, связывающего Z и Z1, на r_1".

(9) IZGLOT:(Z,r)-(Z1,r_1) – "всех слуг узла Z, зависящих от него по отношению r, переподчинить узлу Z1, связав их с Z1 по отношению r_1".

(10) IZSLOT:(Z,r)-(Z1,r_1) – "если у некоторого узла X есть слуга Z, подчиненный X-у по отношению r, то заменить Z новым слугой Z1, подчинив его X-у по отношению r_1".

(11) COGRUZ:(W,W1,Z,r) - "создать дубликат синтаксической группы узла W, назвать вершину созданной группы W1 и соединить Z (хозяина) и W1 (слугу) отношением r".

(12) PERUZOT:(Z,Z1,r)-(Z2,Z1,r_1) – "переподчинить узел

Z1, зависящий от узла Z по отношению r, узлу Z2 по отношению r_1".

Инструкции, изменяющие линейный порядок слов и синтаксических групп во фразе (13-23)

(13) PERLEKRUZ:Z(Z1) – "перенести русскую лексему из узла Z в узел Z1".

(14) PERLEKAUZ:Z(Z1) – то же самое, но применительно к английской лексеме.

(15) IZNOM:Z(Z1) – "изменить номера слов Z и Z1, поменяв их во фразе местами".

(16), (17) PERUZSLEDNOM:Z(Z1) <PERUZPREDNOM:Z(Z1)> – "поставить узел Z во фразе следом за узлом Z1 <перед узлом Z1>".

(18), (19) PERGRSLEDNOM:Z(Z1) <PERGRPREDNOM:Z(Z1)> – "поставить синтаксическую группу узла Z следом за узлом Z1 <перед узлом Z1>".

(20), (21) PERGRMANOM:Z(Z1) <PERGRMINOM:Z(Z1)> – "поставить синтаксическую группу узла Z следом за синтаксической группой узла Z1 <перед этой группой>".

(22), (23) PERUZMANOM:Z(Z1) <PERUZMINOM:Z(Z1)> – "поставить узел Z следом за синтаксической группой узла Z1 <перед этой группой>".

Инструкции, изменяющие словарный состав фразы (24-33)

(24) DOBRUZ:Z(1)> - "добавить к фразе новый узел Z, являющийся русской лексемой 1".

(25) DOBAUZ:Z(1) - то же самое, но применительно к английской лексеме.

(26) ZAMRUZ:Z(1) "заменить лексему, стоящую в узле Z, русской лексемой 1".

(27) ZAMAUZ:Z(1) - то же самое, но применительно к английской лексеме.

(28) PEREVARUZ:Z - "перевести английскую лексему Z на русский язык".

(29) PEREVRAUZ:Z - "перевести русскую лексему Z на английский язык".

(30) STERUZ:Z - "стереть узел Z".

(31) STERGIP:(Z,Z1,r) - "стереть гипотетическую связь с именем r, соединяющую словоформы Z и Z1".

(32) STEROM:Z - "стереть выделенный омоним узла Z".

(33) SOXROM:Z - "сохранить выделенный омоним узла Z, стерев все остальные омонимы этого узла".

Инструкции, обрабатывающие знаки препинания (34-42)

(34) STERPUN:(t_1,...,t_k) - "стереть во фразе знаки препинания t_1,...,t_k".

(35), (36) STERPOUZPUN:Z(t) <STERDOUZPUN:Z(t)> - "стереть знак препинания t, стоящий после узла <перед узлом> Z".

(37), (38) STERPOGUZPUN.Z(t) <STERDOGUZPUN:Z(t)> - "стереть знак препинания t, стоящий после синтаксической группы узла Z <перед этой группой>".

(39), (40) DOBPOUZPUN:Z(t) <DOBDOUZPUN:Z(t)> - "добавить знак препинания t, поставив его после узла Z <перед этим узлом>".

(41), (42) DOBPOGUZPUN·Z(t) <DOBDOGUZPUN:Z(t)> - "добавить знак препинания t, поставив его после синтаксической группы узла Z <перед этой группой>".

Инструкции, работающие с кореферентностью (43-44)

(43) SVUZREF:(Z, Z1) - "связать Z и Z1 кореферентной связью".

(44) IZREF.(Z1,Z2)-(U1,U2) - "заменить в списке кореферентных связей связь между Z1 и Z2 на связь между U1 и U2".

Вспомогательная инструкция (45)

(45) NIHIL: - "ничего не делать".

2.5. Дополнительные указания алгоритму, задаваемые в правилах

2.5.1. Указатель возможной непроективности

В правилах, используемых на этапе синтаксического анализа и описывающих непроективные конструкции, может появиться специальное указание алгоритму о возможной непроективности: NONPR:n. Натуральное число n задает тип непроективности.

2.5.2. Указатель, управляющий обходом структуры

Зона CHECK во всех правилах описывает контекст некоторого фиксированного слова X. Очередное значение X выбирается внешней по отношению к правилам процедурой. Для большой группы правил эта процедура в качестве X должна перебрать

все узлы синтаксической структуры, начиная от ее вершины. Однако преобразования, производимые некоторыми правилами, существенным образом меняют эту структуру. В таких случаях надо подсказать внешней процедуре, с какого узла следует продолжить обход дерева. Запись TAKE:Z означает указание алгоритму: если правило применилось, то обход структуры надо продолжить с узла, являющегося значением переменной Z из условий этого правила. Запись TAKE:1 означает, что обход структуры следует начать с ее вершины.

2.6. Параметры трафаретных правил

В системе ЛП приходится иметь дело с группами правил, совпадающих друг с другом с точностью до некоторых термов (обычно имен лексем). В формальном языке есть удобный аппарат для работы с такими правилами. В сигнатуру языка включен еще один список особого рода переменных — параметров (LR, LR1,..., LA, LA1,..., T, T1,...). Пользователь может ввести в систему правило, которое отличается от обычного лишь тем, что в нем на некоторых аргументных местах (предикатов или инструкций) вместо конкретных термов стоят имена параметров. Такое трафаретное правило, по существу, является схемой правил, обслуживающей некоторую конструкцию. Теперь для получения конкретного правила достаточно указать имя трафаретного правила и список термов для замещения его параметров.

Глава 3
ФОРМАЛЬНАЯ МОДЕЛЬ МОРФОЛОГИИ

3.1. Понятие морфологической структуры

Задача морфологического компонента лингвистического процессора - преобразование входного текста в его морфологическую структуру на этапе анализа и преобразование морфологической структуры в соответствующий ей выходной текст на этапе синтеза [Еськова и др., 1971; Мельчук, 1974]. Под текстом мы далее понимаем последовательность словоформ и знаков препинания, заданных в графической записи. Морфологическая структура (МорфС) текста (в частности, предложения) есть последовательность, элементы которой - МорфС словоформ текста и имена знаков препинания, причем элементы располагаются в том же порядке, что и соответствующие компоненты текста.

Морфологическая структура словоформы есть результат ее лексико-морфологического разбора, т.е. имя соответствующей лексемы, сопровождаемое набором морфологических характеристик. Следует отметить, что по некоторым техническим причинам в морфологических словарях лингвистического процессора содержатся данные о частях речи всех лексем, а для существительных также данные о роде и одушевленности, которые для них являются синтаксическими признаками (см. разд. 6.1.2.3). Эти показатели далее считаются входящими в МорфС словоформы наравне с другими характеристиками. Например, словоформа ***делавшаяся*** имеет морфологическую структуру ДЕЛАТЬ, V, несов, прич, прош, страд, ед, жен, им. Полный перечень характеристик см. в разд. 3.3.

Весьма распространенное явление - морфологическая омонимия, т.е. ситуация, когда словоформа может быть разобрана более чем одним способом. Это могут быть разные формы одной лексемы, например для словоформы ***теории*** мы имеем:

ТЕОРИЯ, S, жен, неод, ед, род;
ТЕОРИЯ, S, жен, неод, ед, дат;
ТЕОРИЯ, S, жен, неод, ед, пр;

ТЕОРИЯ, S, жен, неод, мн, им;
ТЕОРИЯ, S, жен, неод, мн, вин.

В других случаях варианты порождаются разными лексемами - например, для словоформы *спали* мы имеем:

СПАТЬ, V, несов, изъяв, прош, мн;
СПАДАТЬ, V, сов, изъяв, прош, мн;
ПАЛИТЬ, V, сов, пов, ед, 2-л.

Возможна и комбинация этих случаев.

Реже встречается морфологическая синонимия, когда некоторая форма образуется от данной лексемы более чем одним способом. В русском языке есть регулярные случаи синонимии - примерами могут служить сравнительная степень прилагательных и наречий (***сильнее - сильней***) и творительный падеж единственного числа существительных 1-го склонения (***водой - водою***).

При анализе в случае омонимии в МорфС текста включаются все варианты разбора словоформы. Рассмотрим, например, фразу

(1) ***Укажите соединения, которые содержат соли натрия.***

Она получит следующую МорфС (слева указаны номера словоформ во фразе):

МорфС (1)	1	УКАЗЫВАТЬ,	V, сов, пов, мн
	2	СОЕДИНЕНИЕ,	S, сред, неод, ед, род
	2	СОЕДИНЕНИЕ,	S, сред, неод, мн, им
	2	СОЕДИНЕНИЕ,	S, сред, неод, мн, вин
	(после 2-й)		запятая
	3	КОТОРЫЙ,	S, мн, им
	3	КОТОРЫЙ,	S, мн, вин, неод
	4	СОДЕРЖАТЬ,	V, несов, изъяв, непрош, мн, 3-л
	5	СОЛИТЬ,	V, несов, пов, ед, 2-л
	5	СОЛЬ,	S, жен, неод, ед, род
	5	СОЛЬ,	S, жен, неод, ед, дат
	5	СОЛЬ,	S, жен, неод, ед, пр
	5	СОЛЬ,	S, жен, неод, мн, им
	5	СОЛЬ,	S, жен, неод, мн, вин
	6	НАТРИЙ,	S, муж, неод, ед, род
	(после 6-й)		точка.

Отметим, что задачи морфологического анализа и синтеза не являются в точности обратными друг другу, так как МорфС,

подаваемая на вход процедуры синтеза, не должна содержать неоднозначностей, т.е. в ней для каждой словоформы будущего текста задается ровно один вариант лексико-морфологического разбора. При этом процедуре синтеза свойственна и одновариантность на выходе: в случае морфологической синонимии выдается только один, стилистически более предпочтительный вариант.

Из всего сказанного следует, что морфологический анализ и синтез текста фактически сводятся к соответствующим процедурам для отдельных словоформ. Если отвлечься от алгоритмической стороны вопроса, для решения этих задач необходимо и достаточно уметь описывать словоизменение любой лексемы рассматриваемого языка. Другими словами, надо иметь средства, позволяющие для любой лексемы указать множество возможных для нее наборов морфологических характеристик, и каждому набору характеристик из этого множества поставить в соответствие определенную словоформу (при синонимии - несколько словоформ). Эти средства описаны в следующих разделах.

В данной модели морфологии наряду с лексемами вводится еще один тип словарных объектов - так называемые безусловные обороты ("по отношению к", "для того, чтобы", "как бы то ни было" и т.п.). Безусловный оборот есть последовательность слов, имеющих фиксированную форму и следующих друг за другом в фиксированном порядке, которая функционирует в тексте как одно слово. Фразеологические сочетания с изменяемыми или дистантно расположенными компонентами не считаются оборотами и подлежат обычному покомпонентному анализу и синтезу.

3.2. Общие сведения о модели морфологии

Описываемый ниже аппарат формальной морфологии применим к большой группе флективных языков. Поскольку в качестве естественного языка для общения с базами данных был взят русский, аппарат формальной морфологии конкретно применялся к русскому языку. В разд. 3.2 - 3.5 содержится краткое описание русской морфологии, использующее этот аппарат.

Предлагаемая модель русской морфологии удовлетворяет требованиям полноты и адекватности. Это полная модель, описывающая как именное, так и глагольное словоизменение, учи-

тывающая такие сложные явления, как: 1) чередования в основе (ср.: *станок - станк-а*); 2) супплетивизм (*человек - люди*); 3) разные способы видообразования - префиксальный (*#-читать - про-читать*), суффиксальный (*брос-а-ть - брос-и-ть*), чередование основ (*собр-ать - собир-ать*); 4) наличие "добавочных" форм, в частности, падежных форм партитива (*нет сахару*) и местного падежа (*в лесу*); 5) синонимия, ср. "равнозначные" формы (*скорей - скорее*); 6) омонимия (ср. словоформу *стекло*, которая получит несколько альтернативных МорфС: СТЕКЛО, S, сред, неод, ед, им - СТЕКЛО, S, сред, неод, ед, вин - СТЕКАТЬ, V, сов, изъяв, прош, сред, ед).

Четкое отделение содержательного, лингвистического описания от описания алгоритмического позволило нам добиться содержательной адекватности модели русской морфологии, отказаться от выделения единиц, противоречащих (или не вполне соответствующих) лингвистической сущности описываемого объекта.

В задачу морфологического компонента входит установление соответствия между русской словоформой в принятой орфографической записи и МорфС этой словоформы (как уже говорилось, МорфС словоформы - это имя соответствующей лексемы с набором ее морфологических характеристик). Так, в результате работы морфологического компонента устанавливается соответствие между словоформой *делавшаяся* и ее структурой:

делавшаяся ⇔ ДЕЛАТЬ, V, несов, прич, прош, страд, ед, жен, им.

Простейшим способом осуществления подобных преобразований является использование словаря словоформ, где каждой словоформе ставится в соответствие ее структура. Этот путь, возможно, оправданный для языков с бедной морфологией, заведомо неприемлем для русского языка. Достаточно указать, что полная парадигма русского глагола включает 225 словоформ, т.е. лексема ДЕЛАТЬ должна обрабатываться 225 индивидуальными (применимыми только к данной лексеме) правилами типа того, которое было приведено выше.

Нами был принят другой подход, более оправданный и в содержательном, лингвистическом плане, и в плане техническом (см.: [Еськова и др., 1971; Коровина и др., 1977]). Он основан на том, что словоформа разбивается на **сегменты** (бу-

квенные цепочки), соответствующие определенным лингвистически содержательным позициям, и с этими сегментами связываются некоторые морфологические характеристики. Отдельный сегмент вместе с приписанными ему характеристиками называется **морфой**. Для русского языка в словоформах выделяются следующие шесть позиций: 1) приставка (префикс), 2) основа, 3) тема, 4) суффикс, 5) окончание, 6) частица (позиции нумеруются соответственно их месту в словоформе, считая слева направо). Например, словоформа ***делавшаяся*** разбивается на морфы следующим образом:

приставка:	#	(несов)
основа:	***дел***	(V)
тема:	***а***	
суффикс:	***вш***	(прич, прош)
окончание:	***ая***	(ед, жен, им)
частица:	***ся***	(страд)

(в круглых скобках даются морфологические характеристики морф; морфа темы имеет пустой набор характеристик).

Набор характеристик всей словоформы образуется объединением характеристик составляющих ее морф.

Понятие морфы удобно тем, что позволяет описывать целые совокупности словоформ, получаемые независимым варьированием морф разных позиций. Например, заменяя в словоформе ***делавшаяся*** суффикс ***-вш-*** (прич, прош) на ***-ющ-*** (прич, непрош), мы получаем формы причастий настоящего времени; заменяя окончание ***-ая-*** (ед, жен, им) на другие возможные здесь морфы, получаем изменение причастий по роду, числу и падежу; заменяя ***-ся*** (страд) на "#", получаем формы действительного залога (в морфе "#" характеристики нет, так как действительный залог выражается отсутствием характеристики "страд"). С учетом этого приема словоизменение лексемы задается следующим образом: в ее словарной статье указывается одна или несколько основ, а для каждой основы - один или несколько "блоков", каждый из которых содержит списки морф некоторых позиций. Блок порождает словоформы, образованные сочетанием основы с произвольными морфами, указанными в этом блоке (причем требуется, чтобы в словоформе участвовала ровно одна морфа каждой представленной в блоке позиции).

В морфологии русского языка существенную роль играет чередование в основе. Например, существительные КУРОК и СОРОК имеют совпадающие наборы морф-окончаний, однако у большинства форм первого, в отличие от второго, в основе происходит выпадение гласной -О-. Поскольку чередование в основе, по существу, не связано с аффиксальным словоизменением, оно в данной модели описывается автономно, для чего явным образом указывается, каким наборам характеристик соответствуют разные варианты основы. Это можно сделать достаточно экономными средствами (см. разд. 3.4).

В последующих разделах содержится описание основных элементов модели русской морфологии: а) морфологических характеристик (разд. 3.3); б) морфологического словаря (разд. 3.4); в) стандартных объектов, используемых в описании русской морфологии (разд. 3.5).

3.3. Морфологические признаки и характеристики

3.3.1. Вводные замечания

Мы уже видели, что набор характеристик словоформы складывается из наборов характеристик составляющих ее морф. В свою очередь, набор характеристик морфы состоит из одной или нескольких характеристик, представляющих собой значения соответствующих морфологических *признаков (категорий)*. *Так*, в наборе характеристик морфы *-ами* (мн, твор) первая характеристика (мн) - это значение признака "число", а вторая (твор) - значение признака "падеж".

Для русского языка мы выделяем 12 морфологических признаков. Ниже приводится список этих признаков с некоторыми содержательными комментариями, *дающимися в виде примечаний*.

Для каждого признака указываются соответствующие морфологические характеристики и их условные обозначения. Для частей речи, кроме того, приводятся номера признаков, возможных при основах, входящих в данную часть речи. Факультативные признаки (т.е. такие, которые приписываются не всем словоформам, входящим в данную часть речи) даются в круглых скобках. Так, для числительных единственным обязательным признаком считается падеж, а два признака являются факультативными: "одушевленность" (она приписывается только нескольким числительным: ОДИН, ДВА, ТРИ, ЧЕТЫРЕ, ОБА) и "род" (приписывается числительным ОДИН, ДВА, ПОЛТОРА, ОБА).

3.3.2. Список русских морфологических признаков

I. Часть речи

1.	Существительное	S	IV, V
2.	Прилагательное	A	(II - VII)
3.	Глагол	V	(II - V), (VII), VIII, IX, (X - XII)
4.	Наречие	ADV	(VI)
5.	Числительное	NUM	(II), (III), V
6.	Предлог	PR	-
7.	Композит	COM	-
8.	Союз	CONJ	-
9.	Частица	PART	-

Примечание. К композитам мы относим части сложных лексем, ср. композиты ЯРКО-, БЕЛО-, АНГЛО-, НЕМЕЦКО- в примерах: *ярко-зеленый, бело-розовый, англо- и немецко-русский словарь.*

II. Одушевленность

1.	Одушевленное	од
2.	Неодушевленное	неод

III. Род

1.	Мужской	муж
2.	Женский	жен
3.	Средний	сред

IV. Число

1.	Единственное	ед
2.	Множественное	мн

V. Падеж

1.	Именительный	им
2.	Родительный	род
3.	Партитивный	парт
4.	Дательный	дат
5.	Винительный	вин
6.	Творительный	твор
7.	Предложный	пр
8.	Местный	местн

Примечание. Прилагательные, причастия и подавляющее большинство существительных имеют шесть падежей, некоторые существительные - семь или восемь. Дополнительные падежи

партитив и местный) выделяются только в единственном числе некоторых существительных мужского рода. Указания на возможность иметь партитив и/или местный падеж даются в соответствующих статьях комбинаторного словаря (в виде синтаксических признаков ПАРТИТ и ЛОК). Если в ходе синтаксического синтеза характеристики "парт" или "местн" приписаны прилагательному, причастию или такому существительному, которое не имеет этого падежа, то на синтаксическом уровне эти характеристики заменяются соответственно характеристиками "род" и "пр". (Другое возможное решение, согласно которому "парт" и "местн" есть у всех прилагательных, причастий и существительных, но они всегда омонимичны падежам "род" и "пр" соответственно, мы считаем менее предпочтительным хотя бы в силу громоздкости.)

VI. Степень сравнения

1. Сравнительная срав
2. Превосходная прев

Примечание. Особое значение "положительная степень" у признака "степень сравнения" отсутствует. Прилагательные и наречия не сравнительной и не превосходной степени трактуются как не имеющие никакой степени. Данное решение принято отчасти из семантических соображений ("положительной степени" трудно приписать какое-либо самостоятельное означаемое), а отчасти из формальных (эта характеристика реализовалась бы только нулевой буквенной цепочкой и, кроме того, ее пришлось бы включать в МорфС всех наречий и полных прилагательных).

VII. Краткость

1. Краткое кр

Примечание. Мы считаем, что признак "краткость" имеет одно значение - "краткая форма". Тем самым в формах типа *красный, рассмотренная, вымытому* признак "краткость" вообще отсутствует.

VIII. Репрезентация

1. Изъявительное наклонение изъяв
2. Повелительное наклонение пов
3. Инфинитив инф
4. Причастие прич
5. Деепричастие деепр

IX. Вид

1.	Несовершенный	несов
2.	Совершенный	сов

X. Время

1.	Настоящее - будущее (= непрошедшее)	непрош
2.	Прошедшее	прош
3.	Настоящее (для глагола БЫТЬ)	наст

XI. Лицо

1.	Первое лицо	1-л
2.	Второе лицо	2-л
3.	Третье лицо	3-л

XII. Пассивность

1.	Пассивный залог	страд

Примечание. Пары "глагол без частицы -СЯ (-СЬ) / глагол с частицей -СЯ (-СЬ)" объединяются в словаре в одну лексему только в том случае, если они выражают чисто залоговое противопоставление "актив-пассив". Все прочие глаголы с частицей -СЯ, имеющие другие значения - возвратность (БРИТЬСЯ), взаимность (ДРАТЬСЯ) и т.п., считаются самостоятельными лексемами и образуют особые статьи словаря (при этом частице -СЯ (-СЬ) приписывается нулевая характеристика).

3.4. Запись информации в морфологическом словаре

Для записи информации в статьях морфологического словаря применяется аппарат, описанный в работе [Коровина и др., 1977].

Статьи словаря делятся на два типа: 1) статьи, описывающие лексемы (в том числе безусловные обороты) и 2) статьи, описывающие стандартные объекты, т.е. повторяющиеся в разных статьях фрагменты словарной информации.

Каждая статья представляет собой текст, состоящий из **блоков** - отрезков текста без пробелов; блоки отделяются друг от друга произвольным числом пробелов. Длинный блок можно записать в нескольких строках, поместив в каждой из них, кроме последней, специальный знак переноса *, которые указывает, что пробелы между частями блока не должны приниматься во внимание. В любом месте статьи может помещаться комментарий - произвольная последовательность символов заключенная в квадратные скобки.

3.4.1. Лексемы и безусловные обороты

В статье лексемы приводится следующая информация:

1. Номер лексемы (один блок).
2. Имя лексемы (один блок).
3. Основа (один блок).
4. Ограничители (один блок).
5. Признак неокончательности (один блок).
6. Данные о словоизменении (один или несколько блоков).

Статья может содержать несколько групп данных 3-6, соответствующих разным основам. Номер имеет вид пятизначного числа. Имя лексемы, как правило, совпадает с основной формой данной лексемы, к которой справа может быть добавлена десятичная цифра для идентификации одного из значений. Внутри имени может быть поставлен разделитель | (вертикальная черта). Имя лексемы играет в статье вспомогательную роль и может быть опущено.

Основы могут быть двух видов: простые и с чередованием. Блок, задающий простую основу, начинается со служебных символов "осн:", за которыми следует собственно основа - некоторая цепочка символов, не содержащая скобок и двоеточий. Если эта цепочка совпадает с именем лексемы или его частью до знака |, ее можно заменить знаком =.

Основа с чередованием начинается с "осн:", после чего следует X(A)Y либо X(A|B)Y либо X(A|B|C)Y и т.п., где X, Y, A, B, C ... - сегменты (цепочки символов), не содержащие скобок, двоеточий и знака |. Такая основа представляет упорядоченный набор из двух или более простых основ:

```
X(A)Y      =  {XAY, XY};
X(A|B)Y    =  {XAY, XBY};
X(A|B|C)Y  =  {XAY, XBY, XCY}
```

и т.п. Заметим, что A, B, C, ... не могут быть пустыми; если в этой позиции должен стоять пустой сегмент, он заменяется символом #. Таким образом, запись X(A)Y эквивалентна X A|#)Y. На месте X и Y пустые сегменты разрешаются.

Примеры:

```
сос(е)н       (сосна - сосен);
мат(#|ер)     (мать - матери);
креп(ок|ч|к)  (крепок - крепче - крепкий).
```

В блоке, задающем основу, может быть указан список характеристик, приписываемых всем словоформам, образованным от данной основы. Эти характеристики записываются через запятую перед символами "осн:", перед ними ставятся символы "хар:".

Если в основе есть чередование, после нее ставится блок, описывающий распределение вариантов основы по членам парадигмы. Он начинается с символов "чер:", а затем указывается один или несколько ограничителей, каждый из которых имеет вид $X_1/.../X_k$ ($k > 0$), где X_i - список характеристик, разделенных запятыми. В множестве возможных для данной лексемы наборов морфологических характеристик ограничитель выделяет те и только те наборы, которые целиком включают какую-либо из указанных в данном ограничителе групп X_i. Ограничители отделяются друг от друга символом | и соответствуют чередующимся основам (кроме последней), взятым в порядке записи их сегментов внутри скобок. На долю последней основы остаются наборы характеристик, не соответствующие ни одному из указанных ограничителей.

Примеры (справа указана лексема):

сос(е)н	чер:мн,род	(СОСНА);
крас(е)н	чер:кр,муж	(КРАСНЫЙ);
стан(о)к	чер:ед,им/ед,вин	(СТАНОК);
мат(#|ер)	чер:ед,им/ед,вин	(МАТЬ);
креп(ок|ч|к)	чер:кр,муж|срав/прев	(КРЕПКИЙ).

Списки характеристик или их дизъюнкции (в том числе и целые ограничители), встречающиеся во многих статьях, можно объявить стандартными и закрепить за ними определенные номера. Тогда в блоках "чер:" вместо ограничителей или их частей достаточно указывать соответствующие номера, причем в одном блоке могут сочетаться номера и явно заданные характеристики. Примеры:

крас(е)н	чер:6;
стан(о)к	чер:2;
мат(#|ер)	чер:2;
креп(ок|ч|к)	чер:6|срав/прев;
креп(ок|ч|к)	чер:6|10.

Признак неокончательности играет техническую роль. Он показывает, что при морфологическом анализе для словоформ, содержащих данную основу, должна быть сделана попытка их

разбора с более короткими основами (ср.: [Кулагина, 1985, с. 6]). Например, словоформа *паром* может быть образована как от основы ПАРОМ, так и от основы ПАР. Признак неокончательности записывается в виде "ред:A,B", где A, B = 0 или 1, причем 1 означает необходимость рассмотрения более коротких основ. Число A относится к случаю, когда словоформа совпадает с рассматриваемой основой, B - к случаю, когда словоформа длиннее основы. Признак помещается в статью только если A = 1 или B = 1. В русском морфологическом словаре признак неокончательности присутствует приблизительно в 3 % статей.

Данные о словоизменении содержатся в блоках, называемых основными. В основном блоке указывается перечень свободно сочетающихся друг с другом морф различных позиций; разделителем служит запятая. Отдельная морфа записывается в виде выделенного апострофами символьного сегмента, сопровождаемого морфологическими характеристиками. Набор характеристик может быть пустым; возможен и пустой сегмент, изображаемый символом #. Примеры записи морф:

'ее'срав;
'#'ед,им;
'ся';
'ят'изъяв,непрош,мн,3-л.

Позиции морф обозначаются служебными словами "пр:" (приставка), "тм:" (тема), "сф:" (суффикс), "ок:" (окончание), "чс:" (частица). После такого слова указывается одна или несколько морф. Например, основной блок может иметь вид

ок:'#'ед,им,'а'ед,род,'у'ед,дат, *
'#'ед,вин,'ом'ед,твор,'е'ед,пр, *
'ы'мн,им,'ов'мн,род,'ам'мн,дат, *
'ы'мн,вин,'ами'мн,твор,'ах'мн,пр.

Этот блок описывает полное словоизменение существительных типа СТОЛ.

В следующих двух примерах сочетаются морфы разных позиций:

тм:'ова'сов,ок:'ть'инф,'в'деепр,прош;
пр:'#'несов,ок:'ть'инф,чс:'#','ся'страд.

В основном блоке может также присутствовать служебное слово "хар:", после которого указывается одна или несколько характеристик. Эти характеристики считаются относящимися к

основе и действительны для всех вариантов словоизменения, описанных в данном блоке.

Указывая для каждой основы рассматриваемой лексемы один или несколько подобных блоков, можно полностью задать словоизменение лексемы; при этом каждый блок задает совокупность словоформ, образованных всевозможными сочетаниями морф разных позиций, причем в образовании отдельной словоформы участвует ровно одна морфа из каждой позиции. Набор характеристик словоформы формируется объединением характеристик морф, участвующих в ее построении, и характеристик, указанных в группе "хар:" в данном блоке и/или в блоке, задающем основу.

Однако такое явное описание всех форм каждой лексемы было бы чрезвычайно неэкономным, поскольку многие части парадигм и целые парадигмы являются общими для больших классов слов. Чтобы уменьшить объем статей, предусмотрен аппарат сокращений типа макроопределений, суть которого в том, что некоторые стандартные объекты (группы блоков, отдельные блоки или их части) заменяются в описаниях краткими условными обозначениями. Этот аппарат подробно излагается в разд. 3.4.2.

В статье оборота приводится следующая информация:

1. Номер оборота (один блок).
2. Имя оборота (один блок).
3. Текст оборота (один блок).

В статье может быть несколько групп типа 3. Номер оборота - такой же блок, как номер лексемы. Имя оборота - произвольная цепочка символов; пробелы изображаются знаком "_". Этот блок не является обязательным.

Текст оборота начинается со служебного слова "хар:", после которого указывается одна характеристика - часть речи. Затем после запятой ставится служебное слово "об:" и идет собственно текст, где пробелы заменены знаком "_" (в тексте должен быть хотя бы один знак "_"). Если текст совпадает с именем оборота, его можно заменить на "=".

3.4.2. Стандартные объекты

В морфологической информации предусмотрены три основных типа стандартных объектов - трафареты, форматы и стандартные списки, и два вспомогательных - маски и стандартные

ограничители. Стандартный список есть перечень морф одной позиции (через запятую); формат - блок или часть блока; трафарет - блок или группа блоков. Маска есть последовательность нулей и единиц; стандартный ограничитель - набор характеристик или несколько таких наборов, разделенных косой чертой. В трафаретах могут присутствовать параметры, т.е. морфы, конкретный вид которых задается при обращении к данному трафарету.

Стандартные объекты обозначаются в основных блоках служебными словами "т:" (трафарет), "ф:" (формат), "сок:" (стандартный список окончаний), "счс:" (стандартный список частиц). После такого слова ставится десятичное число - номер объекта. После "сок:" может стоять несколько номеров, разделенных запятой. У трафаретов после номера могут в скобках указываться один или два параметра - морфы, перед которыми стоят символы "п1:" и/или "т1:", обозначающие соответственно параметр-приставку и параметр-тему.

Обозначение стандартного объекта, присутствующее в основном блоке, будем называть **обращением** к этому объекту (по аналогии с обращением к подпрограмме). Примеры обращений:

```
сок:39
счс:2
сок:33,37,44
ф:6
т:10
т:61(п1:'по'сов)
т:24(т1:'я'несов)
т:46(п1:'при'сов,т1:'и')
```

Обращение сок:33,37,44 эквивалентно группе обращений сок:33,сок:37,сок:44, что соответствует объединению указанных стандартных списков. В сочетании со стандартными списками окончаний могут использоваться маски. Маска "накладывается" на стандартный список, при этом в списке остаются только те морфы, которым в маске соответствует 1. Так, запись сок:37/8 означает, что в стандартном списке русских окончаний № 37 оставлены морфы, на которые в маске № 8 приходится 1 (здесь важно, что и стандартные списки, и маски линейно упорядочены).

С учетом обращений к стандартным объектам, основной блок в статье лексемы представляет собой либо обращение к трафа-

рету, либо последовательность разделенных запятыми групп следующих типов:

пр: + список морф,
тм: + список морф,
сф: + список морф,
ок: + список морф,
чс: + список морф,
хар: + список характеристик,
счс: + номер,
сок: + список номеров (возможно, с номерами масок, записанными через "/"),
ф: + номер.

Стандартные объекты содержатся в виде отдельных статей в общем морфологическом словаре. В таких статьях первым блоком является пятизначный заголовок (аналог номера в статье лексемы), а затем идет информация, которая должна подставляться вместо обращения к стандартному объекту. Заголовок имеет вид A:XXX, где A - строчная буква, определяющая тип объекта, XXX - трехзначный номер объекта в десятичной записи. Соответствие "буква - тип объекта" следующее:

о - стандартный список окончаний,
м - маска,
ч - стандартный список частиц,
ф - формат,
т - трафарет,
р - стандартный ограничитель.

В стандартных списках указывается один блок - список морф через запятую. В качестве формата может фигурировать любой основной блок или часть основного блока, не содержащие обращений к форматам и трафаретам.

Статья трафарета представляет собой один или несколько произвольных основных блоков, с тем ограничением, что в них не допускаются обращения к трафаретам. Если трафарет имеет параметр-приставку, то в его статье после служебного слова "пр:" может вместо одной из морф стоять обозначение "п1", что означает ссылку на параметр. Аналогичным образом, обозначение "т1" после служебного слова "тм:" означает ссылку на параметр-тему. Фактическими значениями параметров являются морфы, заданные в обращении к трафарету.

Маска есть блок, состоящий из нулей и единиц, разделен-

ных запятыми. Ограничитель - это блок, представляющий дизъюнкцию наборов характеристик, т.е. произвольную последовательность имен характеристик, в которой соседние характеристики разделены знаком "," или "/".

3.4.3. Образцы словарных статей морфологического словаря

Лексемы

```
00024   над   хар:PR,осн:=

00241   допустим|ый   осн:=   ред:1,1
                      ф:4   хар:A,сок:30

00376   много|о   осн:=   ф:38

01097   дел|ать   хар:V,осн:=
           т:26(п1:'с'сов,т1:'а')
           т:45(п1:'с'сов,т1:'а')
           т:64(п1:'с'сов,т1:'а')

01365   плав|ить
        хар:V,осн:=
           пр:'рас'сов,'#'несов,сок:38,46,ф:28
           пр:'#'несов,ф:33,счс:2
           пр:'рас'сов,сок:41
           пр:'#'несов,сок:41,счс:2
           пр:'#'несов,ф:16
           т:45(п1:'рас'сов,т1:'и')
        хар:V,осн:плавл
           пр:'рас'сов,'#'несов,ф:23
           т:69(п1:'рас'сов)

02674   станок   осн:стан(о)к   чер:2   т:9

03238   аналоги|я   осн:=   ред:1,0   т:3

04104   кислород   осн:=   ф:1,сок:6/7

07656   девяност|о   хар:NUM,осн:=
             ок:'о'им,'а'род,'а'дат,'о'вин,      *
                'а'твор,'а'пр

08812   дождев|ой   осн:=   ф:8
```

Безусловные обороты

```
20015   несмотря_на   хар:PR,об:=

20068   в_противном_случае   хар:ADV,об:=
```

Стандартные объекты

м:006 0,0,0,0,0,0,1,1,1,1,1,1

м:011 1,1,1,0,1,1,1,1,1,0,1,1

о:024 'ое'ед,сред,им,'ого'ед,сред,род,'ому'ед,сред,дат, *
'ое'ед,сред,вин,'им'ед,сред,твор,'ом'ед,сред,пр

р:002 ед,им/ед,вин

ф:008 сок:16,19,23,27,хар:A

ф:038 сф:'о',хар:ADV

т:009 ф:1,сок:9

т:064 пр:п1,тм:т1,сф:'нн'прич,прош,страд,сок:13,19,23,27
пр:п1,тм:т1,сф:'н'прич,прош,страд,сок:30

т:106 хар:S,муж,од,сок:6/11,ок:'а'ед,вин,'ов'мн,вин

ч:001 '#','сь'страд

3.5. Стандартные объекты в описании русской морфологии

Русские морфы иногда непосредственно приводятся в соответствующих блоках морфологической словарной статьи. Однако в большинстве случаев в этом нет необходимости, поскольку описанный выше аппарат позволяет применить некоторые естественные способы сокращения записи морфологической информации - стандартные списки, стандартные ограничители, маски, форматы, трафареты (см. разд. 3.4.2). Стандартный ограничитель - это типовое (применимое ко многим лексемам) правило распределения чередующихся основ; маска - правило получения из стандартного списка редуцированной грамматической парадигмы; формат - блок или часть блока, общая для нескольких или многих слов и задаваемая отдельно (в морфологической информации этих слов ставится лишь номер формата, например ф:10); трафарет - блок или совокупность блоков, общая для нескольких или многих слов и задаваемая отдельно.

В результате морфологическое описание подавляющего большинства слов приобретает весьма компактный вид. Так, морфологическая информация лексемы АБОНИРОВАТЬ, задающая 225 словоформ этого глагола с их полными грамматическими характеристиками, выглядит следующим образом:

абонир|овать осн:= т:21.

В настоящее время русский морфологический компонент

включает 47 стандартных списков окончаний, 2 списка частиц, 9 стандартных ограничителей, 42 маски, 42 формата и 93 трафарета.

Лингвистически содержательные объекты - стандартные списки окончаний и частиц, стандартные ограничители - приводятся в этом разделе полностью. Что касается остальных стандартных объектов - масок, форматов, трафаретов, то они имеют более технический характер, и здесь мы сочли возможным ограничиться некоторыми иллюстрациями.

3.5.1. Стандартные списки

Таблица 1

Списки окончаний существительных

№	Единственное число						Множественное число						Пример
	им	род	дат	вин	твор	пр	им	род	дат	вин	твор	пр	
1	А	Ы	Е	У	ОЙ	Е	Ы	#	АМ	Ы	АМИ	АХ	*вод-А*
2	Я	И	Е	Ю	ЕЙ	Е	И	Ь	ЯМ	И	ЯМИ	ЯХ	*пул-Я*
3	Я	И	И	Ю	ЕЙ	И	И	Й	ЯМ	И	ЯМИ	ЯХ	*лини-Я*
4	А	И	Е	У	ОЙ	Е	И	#	АМ	И	АМИ	АХ	*книг-А*
5	Ь	И	И	Ь	ЬЮ	И	И	ЕЙ	ЯМ	И	ЯМИ	ЯХ	*рук-А* *част-Ь*
6	#	А	У	#	ОМ	Е	Ы	ОВ	АМ	Ы	АМИ	АХ	*ротор-#*
7	Ь	Я	Ю	Ь	ЕМ	Е	И	ЕЙ	ЯМ	И	ЯМИ	ЯХ	*кул-Ь*
8	Й	Я	Ю	Й	ЕМ	И	И	ЕВ	ЯМ	И	ЯМИ	ЯХ	*жреби-Й*
9	#	А	У	#	ОМ	Е	И	ОВ	АМ	И	АМИ	АХ	*сток-#*
10	О	А	У	О	ОМ	Е	А	#	АМ	А	АМИ	АХ	*лиц-О*
11	Е	Я	Ю	Е	ЕМ	И	Я	Й	ЯМ	Я	ЯМИ	ЯХ	*войск-О* *мнени-Е*
12	#	#	#	#	#	#	#	#	#	#	#	#	*пальто-#*

Таблица 2

Списки окончаний полных прилагательных и причастий

№	им	род	дат	вин	твор	пр	Примеры
	Единственное число, мужской род						
13	ЫЙ	ОГО	ОМУ	ЫЙ (ОГО)	ЫМ	ОМ	*част-ЫЙ*
14	ИЙ	ЕГО	ЕМУ	ИЙ (ЕГО)	ИМ	ЕМ	(не после Г, К, Х) *давн-ИЙ, хорош-ИЙ*
15	ИЙ	ОГО	ОМУ	ИЙ (ОГО)	ИМ	ОМ	(после Г, К, Х) *тонк-ИЙ*
16	ОЙ	ОГО	ОМУ	ОЙ (ОГО)	ЫМ	ОМ	(не после шип. или Г, К, Х) *пуст-ОЙ*

Окончание таблицы 2

№	им	род	дат	вин	твор	пр	Примеры
17	ОЙ	ОГО	ОМУ	ОЙ (ОГО)	ИМ	ОМ	(после шип. или Г,К,Х) *плох-ОЙ, больш-ОЙ*
18	#	А	У	# (А)	ЫМ	ОМ	*лапласов-#*
	Единственное число, женский род						
19	АЯ	ОЙ	ОЙ	УЮ	ОЙ	ОЙ	(в твор. -ОЙ) *черн-АЯ, больш-АЯ*
20	АЯ	ЕЙ	ЕЙ	УЮ	ЕЙ	ЕЙ	(в твор. -ЕЙ) *хорош-АЯ*
21	ЯЯ	ЕЙ	ЕЙ	ЮЮ	ЕЙ	ЕЙ	*син-ЯЯ*
22	А	ОЙ	ОЙ	У	ОЙ	ОЙ	*лапласов-А*
	Единственное число, средний род						
23	ОЕ	ОГО	ОМУ	ОЕ	ЫМ	ОМ	(в твор. -ЫМ) *черн-ОЕ*
24	ОЕ	ОГО	ОМУ	ОЕ	ИМ	ОМ	(в твор. -ИМ) *тонк-ОЕ, чуж-ОЕ*
25	ЕЕ	ЕГО	ЕМУ	ЕЕ	ИМ	ЕМ	*син-ЕЕ*
26	О	А	У	О	ЫМ	ОМ	*лапласов-О*
	Множественное число						
27	ЫЕ	ЫХ	ЫМ	ЫЕ (ЫХ)	ЫМИ	ЫХ	*прост-ЫЕ*
28	ИЕ	ИХ	ИМ	ИЕ (ИХ)	ИМИ	ИХ	*син-ИЕ, больш-ИЕ*
29	Ы	ЫХ	ЫМ	Ы (ЫХ)	ЫМИ	ЫХ	*лапласов-Ы*

ПРИМЕЧАНИЕ. В скобках указаны окончания одушевленного варианта винительного падежа.

Таблица 3

Списки окончаний кратких прилагательных и причастий

№	муж	жен	сред	мн	Примеры
30	#	А	О	Ы	*прост, прочитан, видим*
31	Ь	Я	Е	И	*син-Ь*
32	#	А	О	И	*строг, хорош*

Ниже приводятся стандартные списки окончаний глаголов (33-46) и числительных (47), а также стандартные списки частиц. Списки даны в их словарной форме.

o:033 'ю'изъяв,непрош,ед,1-л,'ете'изъяв,непрош,мн,2-л
o:034 'ю'изъяв,непрош,ед,1-л,'ите'изъяв,непрош,мн,2-л
o:035 'у'изъяв,непрош,ед,1-л,'ете'изъяв,непрош,мн,2-л
o:036 'у'изъяв,непрош,ед,1-л,'ите'изъяв,непрош,мн,2-л
o:037 'ешь'изъяв,непрош,ед,2-л,'ем'изъяв,непрош,мн,1-л

```
о:038  'ишь'изъяв,непрош,ед,2-л,'им'изъяв,непрош,мн,1-л
о:039  'ет'изъяв,непрош,ед,3-л,'ют'изъяв,непрош,мн,3-л
о:040  'ет'изъяв,непрош,ед,3-л,'ут'изъяв,непрош,мн,3-л
о:041  'ит'изъяв,непрош,ед,3-л,'ят'изъяв,непрош,мн,3-л
о:042  'ит'изъяв,непрош,ед,3-л,'ат'изъяв,непрош,мн,3-л
о:043  'а'ед,жен,'о'ед,сред,'и'мн
о:044  'й',пов,ед,2-л,'йте',пов,мн,2-л
о:045  'и',пов,ед,2-л,'ите',пов,мн,2-л
о:046  'ь',пов,ед,2-л,'ьте',пов,мн,2-л
о:047  'ь'им,'и'род,'и'дат,'ь'вин,'ью'твор,'и'пр
ч:001  '#','сь'страд
ч:002  '#','ся'страд
```

Примечания к стандартным спискам.

1. Парадигмы одушевленных существительных отличаются от соответствующих парадигм неодушевленных существительных формами винительного падежа множественного и (иногда) единственного числа. Однако заводить для одушевленных существительных особые стандартные списки нецелесообразно: они легко описываются с помощью тех же списков 1-12, что и неодушевленные существительные. На стандартный список достаточно лишь наложить маску и дополнительно указать формы винительного падежа (см., например, трафарет т:106 в разд. 3.4.3).

2. Обращает на себя внимание "дробность" стандартных списков окончаний глаголов. Так, список 33 задает всего две формы - первого лица единственного числа и второго лица множественного числа непрошедшего времени. Объединение этих форм с прочими лично-числовыми формами непрошедшего времени в одном стандартном списке мы сочли нежелательным, поскольку они сочетаются с разными частицами. Ср. ***наде-Ю-СЬ, наде-ЕТЕ-СЬ,*** но ***наде-ЕШЬ-СЯ, наде-ЕТ-СЯ, наде-ЕМ-СЯ, наде-ЮТ-СЯ.***

В настоящее время начата реализация другого способа представления форм с частицами -СЯ и -СЬ, содержательно оправданного и технически более целесообразного. Частицы -СЯ и -СЬ рассматриваются как варианты одной частицы -С*, где знак * указывает на чередование Я и Ь (правило распределения задается отдельно). Тем самым необходимость деления парадигмы на дробные списки отпадет.

3.5.2. Стандартные ограничители

```
р:001  ед,им
```

Пример: ***сурок*** (в остальных формах существительного -

основа "сурк": ***сурка, сурку*** и т.д.).

р:002 ед,им/ед,вин

Пример: ***станок*** (в остальных формах - основа "станк": ***станок, станку, ...***),

р:003 мн,род

Пример: ***варежек*** (в остальных формах существительного ВАРЕЖКА - основа "варежк": ***варежка, варежки, ...***).

р:004 мн,род/мн,вин

Пример: ***кошек*** (в остальных формах существительного КОШКА - основа "кошк").

р:005 ед

Пример: ***судно, судна*** и т.д. (в формах множественного числа существительного СУДНО - основа "суд": ***суда, судов, судам, ...***).

р:006 кр,муж

Пример: ***странен*** (в остальных формах прилагательного СТРАННЫЙ - основа "странн").

р:007 мн

Пример: основа "стуль" (***стулья, стульев*** и т.д.); в единственном числе - основа "стул" (***стул, стула, стулу, ...***).

р:008 ед,им/мн,род/мн,вин

Пример: ***турок*** (в остальных формах - основа "турк": ***турка, турку, турки, ...***).

р:009 ед,им/ед,вин/мн,род

Пример: ***валенок*** (в остальных формах - основа "валенк": ***валенка, валенку, ...***).

3.5.3. Примеры форматов и трафаретов

1) ф:007 хар:А,сок:14,20,25,28

Формат задает часть речи (А - прилагательное) и стандартные списки окончаний соответственно мужского, женского, среднего рода единственного числа, а также формы множественного числа, например: ***жгуч-ИЙ, жгуч-ЕГО, ... , жгуч-АЯ, жгуч-ЕЙ, ... , жгуч-ЕЕ, ... , жгуч-ИЕ, ...*** .

2) ф:016 ок:'я'деепр,непрош

Примеры: ***вид-Я, зна-Я.***

3) ф:031 сф:'ющ'прич,непрош,сок:14,20,25,28

Формат задает формы причастий непрошедшего времени с суффиксом 'ющ' (стандартные списки окончаний - те же, что у прилагательных, описываемых форматом 7). Пример: ***кол-ЮЩ-ИЙ,***

кол-ЮЩ-ЕГО, ... , кол-ЮЩ-АЯ, ... , кол-ЮЩ-ИЕ, ...

4) ф:021 сф:'л'изъяв,прош,сок:43

Формат задает часть форм прошедшего времени изъявительного наклонения некоторых глаголов. Пример: ***мерз-Л-А, мерз-Л-О, мерз-Л-И*** (форма мужского рода - ***мерз*** - образуется не с помощью суффикса -Л, а с помощью другого (нулевого) суффикса, и потому не может задаваться тем же форматом, что и другие формы прошедшего времени).

5) т:077 хар:А,сф:'е'срав сф:'айш'прев,ф:7

Трафарет содержит два блока: первый задает форму сравнительной степени прилагательных, образуемую морфой -Е (например, ***крепч-Е***), а второй - формы превосходной степени прилагательных, образуемые с помощью суффикса -АЙШ-; блок включает формат ф:7 (см. выше), задающий сочетающиеся с этим суффиксом окончания: ***крепч-АЙШ-ИЙ, крепч-АЙШ-ЕГО,... , крепч-АЙШ-АЯ, крепч-АЙШ-ЕЙ, ... , крепч-АЙШ-ЕЕ, ... , крепч-АЙШ-ИЕ,....***

6) т:061

пр:п1,'#'несов,сок:33,37,39,44,ок:'ть'инф
пр:'#'несов,ф:31
пр:'#'несов,ф:16
пр:п1,'#'несов,ф:42
пр:п1,'#'несов,ф:40
пр:п1,ок:'в'деепр,прош

Трафарет содержит указания на некоторые стандартные списки окончаний (сок) и форматы (ф), приведенные выше. Он включает шесть блоков, которые задают все формы глаголов типа (ПО)СТРАДАТЬ. Эти глаголы образуют несовершенный вид с помощью нулевого префикса ('#'), а совершенный вид с помощью различных префиксов. Указания на конкретное значение видообразующего префикса даются в статьях соответствующих глаголов, а в записи трафарета префикс задается в виде параметра (см. элемент п1 в четырех из шести блоков, задающих словоизменение). Примеры словоформ, задаваемых трафаретом т:61: ***(ПО)-страда-Ю, (ПО)-страда-ЕШЬ, ... , #-страда-ЮЩ-ИЙ, ... , #-страд-АЯ, (ПО)-страда-Л-#, ...***

3.6. Алгоритмы морфологического анализа и синтеза

В этом разделе дается краткое описание алгоритмов анализа и синтеза словоформы (см.: [Лазурский и др., 1988]).

Все статьи морфологического словаря проходят предварительную трансляцию, в результате которой содержащаяся в них информация приобретает форму, удобную для работы алгоритмов анализа и синтеза. При трансляции статей лексем формируются явные варианты чередующихся основ; если при основе возможна приставка, формируется также основа с "приклеенной" приставкой.

У словоформы, подлежащей анализу, рассматриваются начальные отрезки в порядке убывания их длины. Если некоторый отрезок совпал с имеющейся в словаре основой, делается попытка разобрать словоформу как форму, образованную от данной основы. Для этого перебираются лексемы, имеющие данную основу, и в словарной информации каждой из них рассматриваются соответствующие блоки. Для каждого блока последовательно выполняются две процедуры: сканирование блока и членение словоформы.

Сканирование состоит в том, что блок просматривается от начала к концу и адрес каждой морфы запоминается в поле, соответствующем ее позиции (приставка, тема и т. д.). Когда встречается обращение к стандартному объекту, просмотр продолжается по телу объекта, с возвратом в его конечной точке. В результате возникает удобное развернутое представление блока в виде адресных ссылок на все присутствующие в этом блоке морфы.

Членение осуществляется следующим образом. Если в блоке есть морфы приставок, проверяется, содержит ли данная основа одну из этих приставок. Затем делаются попытки разбить "хвост" словоформы на сегменты, соответствующие указанным в блоке морфам. Это делается простым перебором морф блока, начиная с более левых позиций. Перебор производится полностью; для каждого полученного варианта членения "хвоста" из характеристик основы и морф формируется набор характеристик всей словоформы. В случае чередования в основе проверяется согласованность набора характеристик и рассматриваемого варианта основы.

После того как все возможные варианты разбора словоформы с данной основой построены, процесс анализа, как правило, заканчивается. Исключение составляют случаи, когда при основе указан признак "ред" с 1 в соответствующей позиции (см. разд. 3.4.1). Тогда рассмотрение начальных отрезков

словоформы в порядке убывания длины продолжается, и все делается так, как описано выше.

При синтезе на вход поступает лексема и набор характеристик R искомой словоформы. Для каждого блока выполняется сканирование, а затем в каждой позиции выбирается морфа с наибольшим числом характеристик, обладающая свойством согласованности с R, т. е. не содержащая характеристик, которые не принадлежат R. Характеристики выбранных морф и основы объединяются, и если полученный набор совпадает с R, то конкатенация сегментов этих морф и основы дает искомую словоформу, и процедура синтеза заканчивается. Если же набор характеристик не исчерпывает R, построение искомой словоформы из морф данного блока невозможно. В случае чередования нужный вариант основы выбирается по набору R с учетом информации, заданной в ограничителях.

Корректность описанной процедуры синтеза основана на следующем свойстве реальной морфологической информации: если наборы характеристик двух морф, стоящих в некотором блоке в одной и той же позиции, не вложены друг в друга, то они несовместимы, т. е. не могут одновременно входить в какой-либо возможный для словоформы набор характеристик R.

Глава 4
ФОРМАЛЬНАЯ МОДЕЛЬ СИНТАКСИСА

4.1. Постановка задачи

В данной главе излагается формальная модель русского синтаксиса, лежащая в основе синтаксического компонента нашего лингвистического процессора. В теории "Смысл ↔ Текст", на которую, как уже отмечалось, мы опираемся при создании лингвистической основы процессора, синтаксис понимается как некий универсальный механизм, обеспечивающий соответствие между двумя соседними языковыми уровнями - морфологическим и синтаксическим. Иными словами, синтаксис отвечает за два противоположно направленных типа преобразований: 1) за преобразование морфологической структуры какого-либо естественно-языкового выражения (например, словосочетания, предложения или даже фрагмента текста произвольной длины) в синтаксическую структуру этого выражения и 2) за преобразование синтаксической структуры языкового выражения в его морфологическую структуру.

В действующей модели языка преобразование первого типа принято называть **синтаксическим анализом** (СинтА), а преобразование второго типа - **синтаксическим синтезом** (СинтСз).

Уже отмечалось, что во всех задачах, решаемых лингвистическим процессором, обработка текста производится пофразно. Это означает, в частности, что СинтА отвечает за переход от морфологической структуры русского предложения к его синтаксической структуре, а СинтСз - соответственно от синтаксической структуры предложения к его морфологической структуре. Поскольку морфологическая структура предложения представляет собой последовательность морфологических структур входящих в него словоформ и тем самым является линейным объектом, а синтаксическая структура предложения есть дерево зависимостей, т.е. нелинейный объект, сложность которого существенно выше, понятно, что и сам синтаксический компонент процессора должен быть значительно сложнее, чем другие

его компоненты (причем СинтА сложнее, чем СинтСз). В действительности этот компонент занимает центральное место во всей системе, что обусловливается отнюдь не только техническими причинами, а характером синтаксического уровня самого естественного языка.

Если еще раз обратиться к приведенной в разд. 1.2 блок-схеме ЛП, работающего в режиме анализа, то можно легко убедиться в том, что СинтА является вторым после морфологического анализа звеном цепи, связывающей исходное предложение с его смысловым представлением. На вход этого компонента поступает МорфСпредложения, а его выход - это СинтС предложения, которая, в свою очередь, является входом для следующего этапа работы - семантического анализа.

4.2. Понятие синтаксической структуры предложения

Поскольку понятие синтаксической структуры (СинтС) языкового выражения, и в первую очередь СинтС предложения, по существу, является центральным понятием лингвистического процессора (и, шире, лингвистической модели, лежащей в его основе), представляется разумным рассмотреть его подробнее.

Синтаксической структурой предложения P в данной книге называется размеченное дерево зависимостей, такое, что

1) множество его узлов образуют имена всех лексем, входящих в P;

2) каждая его дуга помечена именем какого-либо синтаксического отношения, специфичного для данного естественного языка.

Идея использования древесной СинтС в качестве промежуточного представления на пути от исходного языкового выражения к его смыслу, коротко говоря, основана на следующем.

Средства, с помощью которых в языке выражается смысл, довольно разнообразны и обладают ярко выраженной национальной спецификой: помимо чисто лексических средств (слов и словосочетаний), смысл передается грамматическими средствами (словоизменительными грамматическими характеристиками типа падежей, родов, глагольных времен и т. п), порядком слов, просодическими средствами (в частности, интонацией), которым на письме соответствует пунктуация, и т. д. Более или менее очевидно, что костяк любого языкового выражения составляет именно лексика, в то время как все другие языко-

вые средства служат в первую очередь для связи между лексическими единицами. Понятно также, что в то время как любое языковое выражение по самой своей природе линейно, связи между его элементами носят далеко не линейный, не одномерный характер (в качестве примера укажем хотя бы на совершенно очевидные связи, имеющиеся в предложении между подлежащим и сказуемым, которые могут стоять далеко друг от друга, определяемым словом и определением к нему и т. п.). Чтобы достаточно адекватно эксплицировать эти связи, необходимо, вообще говоря, использовать многомерные структуры. Однако оперировать "слишком многомерными" структурами при работе с таким и без того сложным объектом, каким является человеческий язык, не так-то просто. Волей-неволей исследователь языка вынужден идти на компромисс, предлагая такие модели для описания языковой действительности, которые, с одной стороны, могли бы достаточно точно и полно отразить моделируемый объект, а с другой - быть обозримыми и легко доступными как для человеческого, так и для "компьютерного" понимания (коль скоро речь идет о модели, предназначенной для создания человеко-машинных систем типа лингвистического процессора).

История теоретической и прикладной лингвистики знает несколько различных моделей, в каждой из которых данный компромисс решается по-своему. Например, в порождающей грамматике Н. Хомского (см., напр.: [Chomsky, 1956, 1957]), которая, по-видимому, была первой из лингвистических моделей, применяемых в компьютерных задачах обработки текстов, используется промежуточное представление фразы в виде системы составляющих (системы синтаксических групп).

Другим результатом этого компромисса является представление синтаксической структуры предложения в виде дерева зависимостей, предложенное независимо Л. Теньером и А. М. Пешковским (см. [Tesnière, 1959; Пешковский, 1956]) и впоследствии получившее развитие в модели "Смысл ↔ Текст".

Заметим для полноты, что принципиально возможен и третий - комбинированный - способ представления синтаксических структур - в виде деревьев зависимостей с составляющими, как это было предложено А. В. Гладким [Гладкий, 1985].

В нашей модели принят второй способ представления синтаксической структуры предложения - дерево зависимостей.

Преимущества дерева как объекта для компьютерной обработки вполне объяснимы: нетрудно увидеть, что дерево - связный граф без циклов – является одной из самых простых нелинейных структур (если вообще не самой простой из них).

Чем же отличается синтаксическая структура некоторого языкового выражения (в частности, предложения) от его морфологической структуры?

Строго говоря, МорфС языкового выражения Р и его СинтС должны быть изоморфны: и тот и другой лингвистические объекты должны содержать информацию, необходимую для извлечения смысла - по существу, инварианта для всех промежуточных представлений Р. Это, в частности, означает, что вся информация, которая передается в МорфС синтаксическими средствами (т. е. грамматическими характеристиками, порядком слов и т. д.), должна сохраняться и при переходе к СинтС. Специфика последней состоит именно в средствах, которые используются для передачи этой информации.

Этим средством являются именованные дуги, соответствующие отношениям синтаксического подчинения, которые связывают слова во фразе. Очевидно, что именем синтаксического отношения можно закодировать ту информацию, которая в МорфС передается чисто согласовательными словоизменительными характеристиками, порядком слов и тому подобными средствами. Иными словами, в СинтС можно не удерживать сведений о тех синтаксических средствах, которые дублируются названиями синтаксических отношений. Именно так строятся СинтС "классической" версии модели "Смысл ↔ Текст". Однако для наших целей по ряду технических соображений оказалось удобным дублировать эту информацию, и наши СинтС являются размеченными и расположенными деревьями зависимостей с полным набором словоизменительных характеристик при узлах.

Приведем в качестве примера СинтС для предложения (1), которая, впрочем, имеет скорее иллюстративный смысл в связи с тем, что словоизменительные характеристики при адъективах нам пришлось опустить. Равным образом здесь, как и во всех других частях книги, нам пришлось отказаться от попытки отразить в приводимых схемах порядок слов.

1) ***Алгоритмы синтаксического анализа и синтеза существенно используют лексикографическую информацию, которая содержится в толково-комбинаторном словаре.***

СинтС предложения (1) имеет вид

4.3. Синтаксические отношения (комментированный перечень)

Из сказанного выше ясно, что информация, передаваемая теми элементами МорфС, которые отсутствуют в СинтС, не должна утратиться она просто передается другими элементами СинтС, для которых в МорфС нет аналога, а именно, синтаксическими отношениями (СинтО). Набор СинтО для русского языка, первоначально предложенный И. А. Мельчуком в книгах [Мельчук, 1964, 1974] (в которых для них использовались, соответственно, термины "отношения непосредственной доминации" и "поверхностно-синтаксические отношения"), был существенно переработан авторами настоящей книги применительно к описываемому лингвистическому процессору. В значительной мере на этот набор отношений оказал влияние список СинтО для английского языка, предложенный в [Mel'čuk, Pertsov, 1987], а также списки СинтО для французского и английского языков, разработанные авторами для систем французско-русского и англо-русского автоматического перевода ЭТАП и ЭТАП-2 (см. [Апресян и др., 1989]; см. также [Иомдин и др., 1975; Иомдин, Перцов, 1975; Саввина, 1976; Урысон, 1981, 1982; Санников, 1989; Иомдин, 1990]).

Ниже следует комментированный перечень русских синтаксических отношений, используемых в настоящее время во всех задачах рассматриваемой системы.

Каждое СинтО соответствует некоторому классу синтаксических конструкций русского языка, имеющих нетривиальные общие свойства. Например, синтаксические связи между членами пары "сказуемое + подлежащее" описываются с помощью единого предикативного СинтО, хотя в русском языке для выражения как сказуемого (вершины предикативного СинтО), так и подлежащего (слуги этого СинтО) используются многообразные средства; ср. *Человек* ←предик— *пришел*; *Мне нравится,* —предик→ *что он умеет владеть собой*; *На эту работу потребуется* —предик→ *от пяти до десяти человек*; *У нас нет* —предик→ *хлеба*; *Прибудет ли самолет вовремя, было неизвестно* (предик) и т. д.

Все СинтО (в текущей версии ЛП их используется 55) являются бинарными и ориентированными.

В приводимом ниже перечне СинтО группируются по признаку их типологической близости; различаются четыре типа СинтО - актантные, атрибутивные, сочинительные и служебные.

Во всех случаях главное слово конструкции обозначается через X, а зависимое - через Y

4 3.1. Актантные СинтО

1. Предикативное СинтО (предик) связывает личную форму глагола или именную часть сказуемого при нулевой связке с именной группой, предложно-именной группой, наречием, инфинитивом, союзом или личной формой глагола придаточного предложения в качестве подлежащего: *Заводы* [Y] *встали* [X], *Люди* [Y] *смертны* [X], *Все три элемента* [Y] *являются* [X] *обязательными*; *Хлеба* [Y] *могло* [X] *не быть*; *Реки* [Y] *не было* [X] *видно*; *Места* [Y] *осталось* [X] *для двоих*; *Их* [Y] *оказалось* [X] *пять*, *Много* [Y] *стульев стояло* [X] *в углу*; *Свыше* [Y] *ста человек явилось* <*явились*> [X] *на субботник*; *С каждого дерева упало* [X] *по* [Y] *груше*; *Дозвониться* [Y] *до него стало* [X] *проблемой*, *Сочтено* [X] *разумным, чтобы* [Y] *сроки работы были откорректированы*; *Может* [X] *оказаться неясным, выполняется* [Y] *ли второе условие* <*почему выполняется* [Y] *второе условие*>, *У меня вызывает* [X] *тревогу, что* Y] *он еще не вернулся*; *Должно* [X] *приниматься во внимание,*

что [Y] *он два года не был в отпуске; Куда он мог* [Y] *деть ключи, никого не интересует* [X]; *Представляется* [X] *сомнительным то* [Y], *чтобы он мог просто отказаться; Меня не очень удивляет* [X] *то* [Y], *что у него все сошлось.*

2. Агентивное СинтО (агент) связывает предикатное слово с его первым аргументом - вершиной именной группы в творительном падеже: *Вопрос рассматривается* ‹*рассмотрен*› [X] *парламентской комиссией* [Y], *прием* [X] *президентом* [Y] *делегации представителей оппозиционных партий.*

3. Квазиагентивное СинтО (квазиагент) связывает предикатное слово (но не глагол в страдательном залоге!) с его первым аргументом - вершиной именной группы в родительном падеже или вершиной предложно-именной группы: *отъезд* [X] *делегации* [Y]; *работа* [X] *компьютера* [Y]; *новые приборы, степень точности* [X] *которых* [Y] *далека от требуемой*; *сообщение* [X] *от* [Y] *нашего корреспондента из Багдада.*

4. Присвязочное СинтО (присвяз) связывает глагол-связку (*быть, казаться, оказываться* и некоторые другие) с именной частью составного сказуемого (фактически вторым синтаксическим актантом связки): *Это мог быть* [X] *кто* [Y] *угодно; Это оказался* [X] *мой знакомый* [Y]; *Результат был* [X] *ошибочный* [Y]; *Чем больше* [Y1] *были* [X1] *сроки, отводимые на выполнение работы, тем ничтожнее* [Y2] *оказывался* [X2] *результат; Жизнь была* [X] *прекрасна* [Y]; *Было* [X] *неизвестно* [Y], *приедет ли он; Дело оказалось* [X] *сложнее* [Y], *чем я думал; Издалека айсберг казался* [X] *огромных размеров* [Y]; *Он показался* [X] *мне больным* [Y]; *Он был* [X] *из* [Y] *дворян; Он был* [X] *как* [Y] *выпад на рапире.*

5. Первое комплетивное СинтО (1-компл) связывает предикатное слово с его вторым синтаксическим актантом (в тех случаях, когда связь не подпадает под определение присвязочного СинтО): *Изыскания* [Y] *продолжает* [X] *небольшая группа ученых; Мы не получали* [X] *такой информации* [Y]; *Мне пришлось сказаться* [X] *больным* [Y]; *продавец* [X] *воздушных шаров* [Y]; *торговец* [X] *овощами* [Y]; *Учитель был рад* [X] *нашим успехам* [Y]; *Я купил* [X] *три яблока* [Y]; *Мы послали* [X] *детям много* [Y] *фруктов; Их рабочие получают* [X] *по* [Y] *пятьсот рублей в месяц; Новый станок обрабатывает* [X] *от* [Y] *70 до 100 деталей в минуту; Мы обратились* [X] *к* [Y] *ним с просьбой; У меня появилось желание* [X] *поскорее закончить*

[Y] *работу*; *Есть мнение* [X], *что* [Y] *наша группа заслуживает поддержки*; *Сознание* [X] *того* [Y], *что ситуация опасна, придавало нам силы*; *Он хочет* [X], *чтобы* [Y] *его оставили в покое*; *Я не знал* [X], *что* [Y] *мой сын целыми днями читает, а не готовится к экзаменам*; *Я не знал* [X], *что (именно) сейчас читает* [Y] *мой сын*; *Спроси* [X], *будет* [Y] *ли сегодня автобус до Москвы.*

6. Второе комплетивное СинтО (2-компл) связывает предикатное слово с его третьим актантом: *прождать* [X] *кого-л. два часа* [Y], *Он вынудил* [X] *нас отказаться* [Y] *от своих замыслов*; *Эти категории мы не считаем* [X] *грамматическими* [Y]; *Назначение* [X] *NN чрезвычайным и полномочным послом* [Y] *в Люксембург произвело эффект разорвавшейся бомбы*; *Дирекция смотрит* [X] *на вашу инициативу как* [Y] *на весьма своевременную и полезную*; *Отправка* [X] *грузов в* [Y] *порт, как обычно, задерживалась*; *Цены на сырую нефть были повышены* [X] *вдвое* [Y]; *Руководитель темы положительно* [Y] *охарактеризовал* [X] *Иванова*; *Он заразил* [X] *меня своим энтузиазмом* [Y], *Это человек, интересный* [X] *для многих своим оригинальным взглядом* [Y] *на жизнь*; *Он предупредил* [X] *нас, что* [Y] *не будет участвовать в этом деле*; *Я вас предупреждал* [X], *чтобы* [Y] *вы не увлекались*; *Он заинтересовал* [X] *меня тем* [Y], *что отказался от публикации своей уже готовой книги*; *Он информировал* [X] *руководство о* [Y] *том, как проекает эксперимент*; *Нас спрашивают* [X], *какие меры будут* [Y] *приняты по этому делу.*

7. Третье комплетивное СинтО (3-компл) связывает предикатное слово с его четвертым актантом: *продавать* [X] *персональные компьютеры по* [Y] *одной - две тысячи долларов за штуку*; *транспортировка* [X] *грузов от причала к* [Y] *контейнерному складу.*

8. Комплетивно-аппозитивное СинтО (компл-аппоз) связывает параметрическое существительное (типа ВЫСОТА, ДЛИНА и т.п.) с беспредложной количественной группой в именительном падеже или с эквивалентным ей наречием или предложной группой, вводимой предлогами ПО, ДО, ОТ: *мачта высотой* [X] *пятьдесят метров* [Y]; *мешки весом* [X] *по* [Y] *пятьдесят килограмм* <*свыше* [Y] *пятидесяти килограмм, до* [Y] *пятидесяти килограмм, от* [Y] *пятидесяти до шестидесяти килограмм*>.

9. Первое несобственно-комплетивное СинтО (1-несобст-

компл) связывает функциональный глагол со смещенным дополнением, представляющим собой первый актант существительного Z – аргумента данного глагола: *Между* [Y] *Англией и Францией шла* [X] *торговая война* [Z]; *разногласия* [Z], *существующие* [X] *между* [Y] *ними.*

10. Второе несобственно-комплетивное СинтО (2-несобст-компл) связывает функциональный глагол со смещенным дополнением, представляющим собой второй актант существительного Z – аргумента данного глагола: *Ответственность* [Z], *которую диспетчер несет* [X] *за* [Y] *все отклонения от режима, весьма велика*; *Помощь* [Z], *оказанная* [X] *первому батальону* [Y], *подоспела вовремя.*

11. Третье несобственно-комплетивное СинтО (3-несобст-компл) связывает функциональный глагол со смещенным дополнением, представляющим собой третий актант существительного Z – аргумента данного глагола: *наказание* [Z], *которому его за* [Y] *это подвергли* [X]; *Он оказал* [X] *на нас этим* [Y] *большое влияние* [Z]

12. Дательно-субъектное СинтО (дат-субъект) связывает существительное в дательном падеже, обозначающее субъект состояния, со словом "категории состояния" или словом типа ДРУГ, СОСЕД, ОТЕЦ и т. п.: *Ему* [Y] *можно* [X] *уйти*; *Ему* [Y] *исполнилось* [X] *бы сегодня 60 лет*; *Мне* [Y] *здесь мелко* [X].

13. Подчинительно-союзное СинтО (подч-союзн) связывает подчинительный союз с вершиной группы сказуемого придаточного предложения: *План будет выполнен, если* [X] *будут* [Y] *в полном объеме поставлены запасные части*; *Рассмотрим треугольник, такой, что* [X] *две его стороны равны* [Y].

14. Инфинитивно-союзное СинтО (инф-союзн) связывает союзы ЕСЛИ, ПРЕЖДЕ ЧЕМ, ЧЕМ, ЧТОБЫ с зависимым инфинитивом: *Чтобы* [X] *выполнить* [Y] *план, необходимо в полном объеме получить запасные части*; *Если* [X] *в качестве неизвестного принять* [Y] *данный параметр, можно продолжить вычисления*; *Прежде чем* [X] *переходить* [Y] *к очередному этапу эксперимента, следует тщательно обдумать полученные результаты.*

15. Сравнительно-союзное СинтО (сравн-союзн) связывает сравнительный союз ЧЕМ или НЕЖЕЛИ с его зависимым: *Новый автомат выполнит эту операцию быстрее, чем* [X] *десять высококвалифицированных рабочих* [Y], *Она чаще бывает задумчива, чем* [X] *беззаботна* [Y].

16. Сравнительное СинтО (сравнит) связывает слово в сравнительной степени или со сравнительным значением с его вторым компаратом: *Ширина отверстия оказалась намного больше* [X] *диаметра* [Y] *стержня*; *Детям старше* [X] *двенадцати лет* [Y] *такого не следует позволять*; *Старый вариант оказался проще и удобнее* [X], *чем <нежели>* [Y] *новый*; *Воздух так* [X] *чист, точно* [Y] *его совсем нет. Нигде не останавливалось столько* [X] *народа, как* [Y] *перед картинной лавочкой* (Н. Гоголь).

17. Предложное СинтО (предл) связывает предлог с постпозитивной или (изредка) препозитивной именной группой: *благодаря* [X] *его настойчивым попыткам* [Y]; *за* [X] *пять суток* [Y]; *к* [X1] *пятому* [Y1] *из* [X2] *столбцов* [Y2], *в* [X1] *черных* [Y1] *и в* [X2] *белых чепчиках* [Y2]; *за* [X1] *пять* [Y1] *или за* [X2] *семь шагов* [Y2], *с* [X1] *пятой* [Y1] *по* [X2] *одиннадцатую страницу* [Y2]; *пятнадцать дней* [Y] *спустя* [X]; *не одних денег* [Y] *ради* [X]; *около* [X] *десяти* [Y], *по* [X] *пять программ* [Y] *в день*; *после* [X] *того* [Y], *как вернетесь*.

18. Элективное СинтО (электив) связывает слово, имеющее показатель или значение выбора каких-либо элементов множества, с предлогом ИЗ: *интереснейшая* [X] *из* [Y] *книг*; *первый* [X] *из* [Y] *докладов*; *самый эффективный* [X] *из* [Y] *этих способов*.

4.3.2. Атрибутивные СинтО

19. (Собственно) атрибутивное СинтО (атриб) связывает существительное или прилагательное с его несогласованным определением: *стол* [X] *Петра* [Y]; *проблема* [X] *необычайной важности* [Y]; *обработка* [X] *давлением* [Y], *дети* [X] *моложе* [Y] *16 лет*; *символ* [X] *справа* [Y], *Олимпиада* [X] – *88* [Y]; *пятое* [X] *февраля* [Y]; *хороший* [X] *по* [Y] *качеству*; *крупнейший* [X] *в* [Y] *Европе*; *теоретически* [Y] *неубедительный* [X].

20. Определительное СинтО (опред) связывает существительное или прилагательное X с прилагательным Y, полностью или частично согласованным с X по роду, числу, падежу и одушевленности: *непосредственная* [Y] *задача* [X], *красный* [Y] *и зеленый шары* [X], *три странных* [Y] *субъекта* [X], *кое-что* [X] *весьма интересное* [Y], *самый* [Y1] *сильный* [X1,Y2]

шахматист [X2] *мира; в каком* [Y] *бы тяжелом* [X] *положении мы ни находились; каждый* [Y1] *третий* [X1,Y2] *житель* [X2].

21. Описательно-определительное СинтО (оп-опред) связывает существительное с обособленным определением: *Третье (на этот раз отрицательное* [Y1]) *решение* [X1] *получится при разложении* [X2], *описанном* [Y2] *в параграфе 4; Сообщения* [X] – *цифровые* [Y], *текстовые, графические – можно хранить в такой системе долгое время.*

22. Аппроксимативно-порядковое СинтО (аппрокс-порядк) связывает существительное с постпозитивным определением к нему, выраженным порядковым прилагательным, образуя конструкцию со значением приблизительного порядка: *день* [X] ←предл *на пятый* [Y], *попытки* [X] *с пятнадцатой* [Y]. (Ср. ниже аппроксимативно-количественное СинтО, п. 42.)

23. Релятивное СинтО (релят) связывает вершину именной группы главного предложения с вершиной группы сказуемого придаточного определительного предложения: *Перечислите фамилии служащих* [X] *отдела сбыта, которые зарабатывают* [Y] *больше, чем в среднем по фирме; Он жил на довольно грязной улице* [X], *которых в Москве было* [Y] *сколько угодно; обыкновенный осинник* [X], *какие часто можно* [Y] *встретить в этих местах; Дома* [X], *что стоят* [Y] *в глубине двора, будут снесены; Вот тот человек* [X], *с кем <с которым> меня неожиданно столкнул* [Y] *случай; Лучшее* [X], *что у него было* [Y], *он отдавал людям; Всякий* [X], *кто осмелится* [Y] *нарушить приказ, подлежит аресту; Все новое* [X], *что есть* [Y] *в этой статье, можно изложить на двух страничках; там* [X], *где еще не ступала* [Y] *нога человека; теперь* [X], *когда все уже решено* [Y]; *Маралы* [X], *как называют* [Y] *этих оленей на севере, занесены в Красную книгу; Вот вкратце то* [X], *о чем я прочел* [Y] *в его письме.*

24. Обстоятельственное СинтО (обст) связывает глагол в качестве вершинного элемента с обстоятельством, выраженным наречием, существительным в творительном, родительном (так называемый "родительный даты") или (изредка) винительном падеже, порядковым прилагательным, деепричастием, инфинитивом, предложно-именной группой, количественной беспредложной группой, придаточным, вводимым союзом, или союзной группой: *Два сопротивления подключены* [X] *параллельно* [Y], *Несколько* [Y] *раз посылались* [X] *рекламации; Система, полу-*

чая [Y] *информацию о приеме заказов, регулирует* [X] *выпуск запасных частей в требуемом количестве и ассортименте; идти* [X] *полем* [Y]; *понимать* [X] *умом* [Y]; *плавать* [X] *брассом* [Y]; *приехать* [X] *поздней осенью* [Y]; *прибыть* [X] *самолетом* [Y]; *работать* [X] *топором* [Y]; *приступить* [X] *к работе двадцатого числа* [Y] <*двадцатого* [Y] *сентября*>; *Уходи* [X] *сию же секунду* [Y]; *Каждые тысячу* [Y] *лет на Земле происходит* [X] *крупная геологическая катастрофа; идти* [X] *милю* [Y]; *набит* [X] *до* [Y] *отказа; Студента избили* [X] *прямо на* [Y] *улице; Юра спокойно отправился* [X] *гулять* [Y]; *Если* [Y] *теорема верна, то будет* [X] *справедливо следующее утверждение.*

25. Обстоятельственно-тавтологическое СинтО (обст-тавт) связывает глагол в качестве вершины с семантически родственным ему существительным в творительном падеже, обозначающим действие, инструмент действия, способ действия и т.п.: *смотреть* [X] *изучающим взглядом* [Y], *жить* [X] *беспечной жизнью* [Y]; *говорить* [X] *грудным голосом* [Y].

26. Субъектно-обстоятельственное СинтО (суб-обст) связывает глагол в качестве вершины с именной группой в творительном падеже, имеющей значение совокупности или вместилища и характеризующей субъект данного глагола: *Птицы селились* [X] *на озере целыми стаями* [Y]; *Студенты толпами* [Y] *ходили* [X] *по улицам.*

27. Объектно-обстоятельственное СинтО (об-обст) связывает глагол в качестве вершины с именной группой в творительном падеже, имеющей значение совокупности или вместилища и характеризующей объект данного глагола: *отправлять* [X] *книги ящиками* [Y]; *Нас повсюду водили* [X] *толпой* [Y].

28. Длительное СинтО (длительн) связывает глагол, не имеющий валентности "длительности" и являющийся вершиной, с обстоятельством - именной группой в винительном падеже, ее синтаксическим эквивалентом или наречием со значением длительности, обозначающими временной интервал: *Аккумулятор работает* [X] *всего год* [Y] <*две* [Y] *пятых положенного времени*> (ср. трактовку аналогичного зависимого при глаголе ПРОРАБОТАТЬ, у которого есть длительная валентность: 1-компл

проработал всего год); *Уже третий час* [Y] *электричество включено* [X]; *Сколько* [Y] *часов идет* [X] *поезд до Варша-*

вы?; Первую [Y] *из этих трех недель он работал* [X] *на стройке; Маленькие дети должны спать* [X] *по* [Y] *десять - двенадцать часов в сутки <не менее* [Y], *чем по десять - двенадцать часов в сутки>; Он долго* [Y] *не приходил* [X].

29. Кратно-длительное СинтО (кратно-длительн) связывает глагол в качестве вершины с обстоятельством - существительным в творительном падеже множественного числа, обозначающим отрезок времени: *Старые грузовики в гараже простаивают* [X] *неделями* [Y] *и месяцами; Яд лягушки Коха годами* [Y] *остается* [X] *активным; Он часами <целыми днями>* [Y] *просиживал* [X] *над отчетами.*

30. Субъектно-копредикативное СинтО (суб-копр) связывает глагол в качестве вершины с копредикативным членом, характеризующим субъект этого глагола и выраженным именной группой в творительном или (реже) именительном падеже или предложно-именной группой: *Он вернулся* [X] *из экспедиции помолодевшим* [Y], *Он и его брат уехали* [X] *победителями* [Y]; *Они ввалились* [X] *в комнату пьяные* [Y]; *Мальчик пришел* [X] *в класс с* [Y] *забинтованной головой.*

31. Объектно-копредикативное СинтО (об-копр) связывает глагол в качестве вершины с копредикативным членом, характеризующим объект этого глагола и выраженным именной группой в творительном или винительном падеже или предложно-именной группой: *Его доставили* [X] *в больницу умирающим* [Y] *<без* [Y] *сознания>; Я помню* [X] *его молодого* [Y] *и здорового.*

32. Количественно-копредикативное СинтО (колич-копред) связывает глагол с количественной группой или ее эквивалентом в роли копредикативного члена: *Книг привезли* [X] *целый ящик* [Y]; *Журналов имеется* [X] *два комплекта* [Y]; *Таких случаев мы различаем* [X] *ровно три* [Y] *<свыше* [Y] *десяти>*.

33. Неактантно-комплетивное СинтО (неакт-компл) связывает глагол, способный иметь неактантное (по существу - смещенное) дополнение, реализующее некоторую синтаксическую валентность слова Z, выступающего при данном глаголе в качестве настоящего дополнения, с существительным в дательном падеже, предложно-именной группой с направительным предлогом или эквивалентным ей направительным наречием: *войти* [X] *к* [Y] *кому-л. в комнату* [Z]; *подняться* [X] *к* [Y] *нему на чердак* [Z]; *посмотрел* [X] *тигру* [Y] *в глаза* [Z]. С помощью

неактантно-комплетивного СинтО оформляются также некоторые конструкции с dativus eticus и dativus commodi, такие, как *позвонить* [X] *ему* [Y] *в Москву*; *крикнул* [X] *мне* [Y] *наверх*, *купил* [X] *ему* [Y] *книги*; *решил* [X] *сыну* [Y] *задачу*.

34. Ограничительное СинтО (огранич) связывает словоформу с частицей или ограничительным наречием: *Мы ведь* [Y1] *еще* [X1,Y2] *даже* [X2,Y3] *не* [X3,Y4] *знакомы* [X4]; *Я* [X] *тоже* [Y] *хочу с вами поговорить*; *Он только* [Y1] *этого* [X1] *и* [Y2] *хочет* [Y2]; *Чем богаче формальный язык, тем* [Y] *мощнее* [Y] *лингвистический процессор*; *Часто* [X] *ли* [Y] *у вас отключают энергию?*; *Первый* [X] *же* [Y] *трансформатор вышел из строя через 2 дня*; *Программа* [X1]*-то* [Y1] *ошибок не содержит, а вот* [Y2] *компьютер* [X2] *сбоит*; *Ни* [Y1] *шагу* [X1] *назад мы не* [Y2] *сделаем* [X2]

35. Количественно-ограничительное СинтО (колич-огран) связывает прилагательное или наречие в сравнительной степени с предлогами В1 или НА1, вводящими количественную группу, или с аналогичными по смыслу наречиями и существительными: *в* [Y] *три раза больше* [X]; *на* [Y] *три метра длиннее* [X]; *вдвое* [Y] *мощнее* [X], *гораздо* <*намного*> [Y] *производительнее* [X]; *часом* [Y] *раньше* [X].

36. Вводное СинтО (вводн) связывает вершину глагольной группы главного предложения в качестве хозяина с вершиной вводной конструкции, вводного оборота или вводным словом в качестве зависимого элемента: *О последствиях, разумеется* [Y], *не подумали* [X], *Эта проблема, конечно* [Y] *же*, *существует* [X]; *Десять сеансов, утверждает* [Y] *профессор Ямада, дают* [X] *стопроцентный успех*; *Последовательность аминокислот в белке, это было* [Y] *совершенно точно доказано, задается* [X] *последовательностью триплетов в ДНК*; *Дом, можно* [Y] *считать, хороший* [X]; *Правда* [Y], *в этом алгоритме не удается* [X] *избежать полного перебора гипотез*; *Он, случается* [Y], *опаздывает* [X] *на лекции*.

37. Изъяснительное СинтО (изъясн) связывает вершину группы сказуемого главного предложения с вершиной группы сказуемого придаточного предложения, вводимого относительными союзными словами ЧТО и ПОЧЕМУ: *Комплекс программ* [X] *был завершен, что позволило* [Y] *начать эксперименты*; *Комплекс автоматически меняет* [X] *свою конфигурацию, в результате чего* <*почему*> *система и сохраняет* [Y] *стабильность*.

38. Аппозитивное СинтО (аппоз) связывает существительное-хозяина с приложением в качестве зависимого члена: *страны* [X] – *участники* [Y] *хельсинкского процесса*; *антенна* [X] – *зонд* [Y], *кандидат* [X1] *исторических наук доцент* [Y1,X2] *Иванов* [Y2]; *в городе* [X] *Москве* [Y], *в марте* [X] *месяце* [Y].

39. Нумеративно-аппозитивное СинтО (нум-аппоз) связывает существительное, обозначающее элемент такого множества предметов, в котором каждый предмет естественно нумеруется, с числительным, задающим некоторый номер: *комната* [X] *триста двадцать семь* [Y], *параграф* [X] *25* [Y], *пункт* [X] *восемь* [Y] и т. п.

40. Композитное СинтО (композ) связывает прилагательное в качестве вершины с композитным элементом, по сути дела формируя составное слово: *англо* [Y] *–русский* [X] *словарь*, *черно* [Y] *–белая* [X] *фотография*; *один* [Y] *другого* [X]

41. Количественное СинтО (количест) связывает существительное или его синтаксический эквивалент в качестве вершины с зависимым числительным: *пять* [Y] *приборов* [X], *двадцать* ←колич–вспом— *одни* [Y] *сутки* [X]; *мы* [X] *оба* [Y], *к нам* [X] *обоим* [Y]

42. Аппроксимативно-количественное СинтО (аппрокс-колич) связывает существительное с постпозитивным зависимым числительным, образуя количественную группу со значением приблизительности: *заказов* [X] *десять* [Y], *человек* [X] *пятьдесят* [Y]; *Ему было лет* [X] ←предл— *под сорок* [Y]

43. Распределительное СинтО (распред) связывает существительное со значением ‘денежная единица’ и именную группу в именительном падеже, обозначающую единицу измерения, либо существительное со значением ‘единица измерения’ и предложно-именную группу, вводимую предлогами В1, НА1, ЗА1: *На мировом рынке уголь стоит до 150 долларов* [X] *тонна* [Y], *со скоростью 100 километров* [X] *в* [Y] *час*; *9 метров* [X] *за* [Y] *секунду*; *пять килограммов* [X] *на* [Y] *погонный метр*.

4.3.3. Сочинительные СинтО

Все сочинительные СинтО в нашей системе представлены как ориентированные, подчинительные. Для определенности принято решение проводить синтаксические связи во всех типах сочинительных синтаксических конструкциях слева направо.

44. Сочинительное СинтО (сочин) связывает первый (очередной) член сочиненной группы со вторым (последующим). В случае союзного сочинения в качестве второго (зависимого) члена сочинительной конструкции выступает союз: *В дыму* [X] —сочин→ *и* [Y] *огне метались люди; Мы купили яблоки* X1], —сочин→ *груши* [Y1,X2] —сочин→ *и* [Y2] *другие фрукты; Текла снеговая* [X1], —сочин→ *чистая* [Y1,X2], —сочин→ *пахучая* [Y2] *вода* (А.Толстой); *точа* [X] *алмаз, дробя* [Y] *гранит* (Н.Гумилев); *Не слыхали* [X1], *не видали* .Y1,X2] *и* [Y2] *не знаем ничего* (Д.Хармс); *фото-* [X1], *кино-* 'Y1,X2] *и* [Y2] *радиоаппаратура; кто* [X1], *куда* [Y1,X2] *и* .Y2] *на какой срок командирован; по* [X] *оврагам и* [Y] *по скатам; Кого* [X] *и* [Y] *о чем вы спрашивали?*

45. Сочинительно-союзное СинтО (соч-союзн) связывает сочинительный союз со следующим за ним членом сочиненной группы: *в дыму* —сочин→ *и* [X] —соч-союзн→ *огне* [Y]; *без шапки* —соч-союзн→ *и* [X] *налегке* [Y], *Первая звезда блеснула надо мной и* [X] *упала* [Y] *в тучи; Куда и* [X] *с* [Y] *кем он пошел?*

46. Сентенциально-сочинительное СинтО (сент-соч) связывает вершины двух однородных предложений: *Дверь скрипнула* —сент-соч→ X], *и* [Y] —соч-союзн→ *вошел хозяин* (Л.Толстой); *Чернели* [X] *пятна* —сент-соч→ *от костров, зола белела* [Y], *кости* (В.Хлебников).

47. Кратное СинтО (кратн) связывает два одинаково оформленных существительных, числительных или прилагательных, разделенных предлогом НА1 или знаком препинания типа дефиса, тире, двоеточия и т.п.: *счет 3* [X]*:4* [Y], *при счете* ять [X] – *шесть* [Y], *размер 5* [X] × *6* [Y], *емкость экрана - 384* [X] *на* [Y] *256 точек; матч Каспаров* [X] – *Карпов* [Y], *обязательные* [X] */ факультативные* [Y] *позиции.*

4.3.4. Служебные СинтО

48. Аналитическое СинтО (аналит) связывает элементы аналитических глагольных форм, при этом вершиной считается изменяемая форма: *Работа будет* [X] *продолжаться* [Y] (аналитическое будущее время); *О какой бы* [Y] *книге мы ни говорили* [X] (аналитическое сослагательное наклонение).

49. Пассивно-аналитическое СинтО (пасс-анал) связывает вспомогательный глагол БЫТЬ в качестве вершины со страдательным причастием совершенного вида в качестве зависимого элемента: *был* [X] *разбит* [Y], *будучи* [X] *брошен* [Y], *будет* [X] *выполнен* [Y]

50. Количественно-вспомогательное СинтО (колич-вспом) связывает числительные и порядковые прилагательные с препозитивным числительным в составе сложных имен чисел: *страница триста* [Y1] *сорок* [X1,Y2] *один* [X2], *триста* [Y1] *сорок* [X1,Y2] *первая* [X2] *страница.*

51. Соотносительное СинтО (соотнос) связывает первый и второй элементы разрывных парных союзов, предлогов и частиц: *Только <едва>* [X] *мы вошли, как* [Y] *дверь распахнулась; Если* [X] *техника будет поставлена до нового года, то* [Y] *заказ может быть выполнен в срок; Чем* [X] *лучше линия будет работать, тем* [Y] *скорее мы выполним план; Поезд туда идет от* [X] *семи до* [Y] *девяти часов; Как* [X] *толстоклювые, так* [Y] *и тонкоклювые кайры предпочитают перемещаться, не опираясь на цевку; Он никогда не читает ни* [Y] *книг, ни* [X] *газет.*

52. Эксплетивное СинтО (эксплет) связывает слова типа ТО, ТОТ в качестве вершины с теми членами предложения, которые ими замещаются: *предчувствие того* [X], *что* [Y] *дело кончится плохо; перед тем* [X] *как* [Y] *совершенно упасть духом; с того* [X] *дня, как* [Y] *он уехал; Я обратился к вам только потому* [X], *что* [Y] *других специалистов в Москве не оказалось.*

53. Разъяснительное СинтО (разъяснит) связывает слово, имеющее родовое или кванторное значение, с обозначением видов данного рода или элементов множества: *В продаже было все* [X] – *колбасы* [Y], *сыры, кондитерские изделия; В продаже нет самых обычных молочных продуктов* [X] *молока* [Y], *масла, творога, сметаны.*

54. Примыкательное СинтО (примыкат) используется для описания внешних связей слов, заключенных между двумя скобками. Принимается, что внутрь скобки входит ровно одна связь, вершиной которой является слово, стоящее левее группы слов, находящихся в скобках (обычно непосредственно перед левой скобкой); зависимым членом примыкательного СинтО является главный член скобочной группы слов: *Чехословацкая*

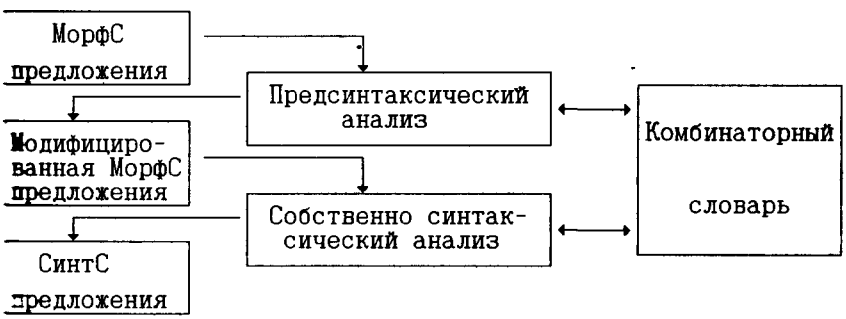
Федеративная Республика [X] *(ЧССР* [Y]); *Директор* [X] *(он выступал* [Y] *первым) сказал, что институт стал отставать от мирового уровня.*

55. Неидентифицирующее СинтО (неидент) используется для введения в СинтС предложения содержащихся в нем неопознанных текстовых единиц (слов, отсутствующих в словаре, формул, неалфавитных символов и т. д.) в тех случаях, когда синтаксическая позиция таких единиц не дает оснований сформировать нормальную синтаксическую связь с помощью какого-либо из полноценных СинтО: *рейс Денвер* [X1] – неидент → *Колорадо* [Y1,X2] *Спрингс* [Y2] (названия соответствующих городов не включены в рабочий словарь ЛП).

4.4. Принципиальная схема синтаксического анализа

Теперь, когда мы имеем достаточно подробное представление о МорфС и СинтС, т. е. о входе и выходе синтаксического компонента ЛП, и лингвистических средствах, используемых в обеих этих структурах, стоит несколько детальнее, чем это было сделано в разд. 4.1, рассмотреть принципиальную схему синтаксического анализа (т. е. с большим увеличением взглянуть на тот фрагмент блок-схемы ЛП, приведенной в гл. 1, который соответствует его синтаксическому компоненту). Подчеркнем, что в этом разделе мы описываем процесс синтаксического анализа с точки зрения лингвиста, имея в виду отразить ту роль, которую играют в этом процессе различные лингвистические блоки системы; математически строгий алгоритм синтаксического анализа, положенный в основу действующего программного комплекса, излагается в разд. 4.8.

Соответствующий "увеличенный" фрагмент блок-схемы ЛП выглядит следующим образом:



Как явствует из рисунка, синтаксический анализирующий

механизм, который для нас до сих пор был единым, на самом деле распадается на два блока: 1) блок предсинтаксического анализа (имеющий, как будет показано, вспомогательный характер) и 2) блок синтаксического анализа в собственном смысле.

Ниже эти два блока будут рассмотрены поочередно.

4.5. Предсинтаксический анализ

Необходимость введения вспомогательного блока предсинтаксического анализа (ПредсинтА) обусловлена следующими причинами.

Говоря о том, что результатом работы морфологического анализирующего компонента ЛП является МорфС предложения, мы в действительности несколько упрощали картину. Дело в том, что практически любое русское предложение содержит словоформы, являющиеся лексически и/или грамматически омонимичными, и чем больше в предложении таких словоформ, тем в нем выше коэффициент омонимии - отношение числа омонимов к числу словоформ, составляющих предложение. Разрешить лексико-грамматическую омонимию средствами морфологического анализа, который, как было показано в гл. 3, работает только на пространстве одного слова (или безусловного оборота) и не может обращаться даже к ближайшему контексту, в принципе невозможно. Из этого факта со всей определенностью следует, что морфологический анализ не может для произвольного русского предложения построить соответствующую ему правильную МорфС (или, в случае неоднозначности, несколько МорфС). Тем самым мы оказываемся перед необходимостью следующего выбора: либо поочередно разбирать всевозможные последовательности МорфС словоформ, составляющих предложение (а таковых, при высоком коэффициенте омонимии, может оказаться достаточно, чтобы вызвать комбинаторный взрыв), либо ввести некоторый конструкт, который объединял бы такие последовательности в пределах одного объекта. Естественно, мы предпочли второй путь.

4.5.1. Комбинированная морфологическая структура предложения

Проиллюстрируем сказанное простым примером. Русское предложение

(2) ***Домик стоит на лесной поляне***

в целом не является омонимичным и допускает ровно одно осмысление. Соответственно, у (2) должна быть ровно одна МорфС. Она имеет следующий вид (как и выше, кортеж грамматических характеристик записывается в виде подстрочных индексов):

2') ДОМИК$_{ед,им}$ СТОЯТЬ$_{изъяв,непрош,несов,3-л,ед}$ НА2 ЛЕСНОЙ$_{жен,ед,пр}$ ПОЛЯНА$_{ед,пр}$.

Между тем каждая из словоформ, составляющих (2), имеет либо лексические, либо грамматические омонимы. В частности, словоформа *домик* имеет два грамматических омонима (им. и вин. падежи); словоформа *стоит* - по крайней мере два лексических омонима (соответствующие лексемам СТОЯТЬ и СТОИТЬ); словоформа *на* - тоже два лексических омонима (в нашем комбинаторном словаре различаются предлоги НА1, управляющий винительным падежом, и НА2, управляющий предложным падежом); словоформа *лесной* - целых шесть грамматических омонимов (им. и вин. падежи муж. рода, род., дат., твор. и предл. падежи жен. рода); наконец, словоформа *поляне* - два грамматических омонима (дат. и предл. падежи). Делением общего числа омонимов предложения (2) на число слов этого предложения мы можем определить, что коэффициент его омонимичности составляет 2,8, а перемножением числа омонимов найдем, что для этого предложения, в случае, если бы мы приняли первый подход, пришлось бы рассмотреть 96 последовательностей МорфС словоформ - кандидатов в МорфС предложения, и все эти последовательности попытаться пропустить через блок синтаксического анализа.

Сделанный нами выбор исключает такую перспективу. В результате морфологического анализа предложения (2) мы в действительности получаем ровно один объект (назовем его комбинированной МорфС предложения): это кортеж, состоящий из множеств, образованных всеми омонимами каждой словоформы данного предложения; ср.

2а') {ДОМИК$_{ед,им}$ / ДОМИК$_{ед,вин}$} {СТОЯТЬ$_{изъяв,непрош,несов,3-л,ед}$/ СТОИТЬ$_{изъяв,непрош,несов,3-л,ед}$} {НА1/НА2} {ЛЕСНОЙ$_{муж,ед,им}$ / ЛЕСНОЙ$_{муж,ед,вин,неод}$ / ЛЕСНОЙ$_{жен,ед,род}$/ЛЕСНОЙ$_{жен,ед,дат}$/ЛЕСНОЙ$_{жен,ед,твор}$/ ЛЕСНОЙ$_{жен,ед,пр}$} {ПОЛЯНА$_{ед,дат}$/ПОЛЯНА$_{ед,пр}$}.

Аналогичным образом, для предложения (1) на выходе морфологического анализатора получается в реальности не МорфС (1'), а кортеж

(1а') {АЛГОРИТМ$_{мн,им}$ / АЛГОРИТМ$_{мн,вин}$} {СИНТАКСИЧЕСКИЙ$_{муж,ед,род}$ / СИНТАКСИЧЕСКИЙ$_{муж,ед,вин,од}$ / СИНТАКСИЧЕСКИЙ$_{сред,ед,род}$} АНАЛИЗ$_{ед,род}$ {И1/И2} СИНТЕЗ$_{ед,род}$ {СУЩЕСТВЕННО/СУЩЕСТВЕННЫЙ$_{кр,сред}$} {ИСПОЛЬЗОВАТЬ$_{изъяв,непрош,несов,3-л,мн}$ / ИСПОЛЬЗОВАТЬ$_{изъяв,непрош,сов,3-л,мн}$} ЛЕКСИКОГРАФИЧЕСКИЙ$_{жен,ед,вин}$ ИНФОРМАЦИЯ$_{ед,вин}$, КОТОРЫЙ$_{жен,ед,им}$ {СОДЕРЖАТЬСЯ$_{изъяв,непрош,несов,3-л,ед}$ / СОДЕРЖАТЬ$_{изъяв,непрош,несов,страд,3-л,ед}$} {В1/В2/В3} ТОЛКОВО - КОМБИНАТОРНЫЙ$_{муж,ед,пр}$ СЛОВАРЬ$_{ед,пр}$.

Именно комбинированные МорфС типа (1а') и (2а') поступают на вход синтаксического анализирующего компонента ЛП.

4.5.2. Назначение блока предсинтаксического анализа

Понятно, что задача построения правильной СинтС для объектов типа (1а') и (2а') (а этим объектам, т.е. предложениям (1) и (2), соответствуют единственные СинтС) существенно упростится, если мы сумеем предварительно упростить сами эти объекты, т.е., говоря иными словами, сократить коэффициент омонимичности предложений. Оказывается, что во многих случаях это достаточно легко сделать, рассмотрев ближайшее окружение омонимичных словоформ. Сказанное можно проиллюстрировать уже приводившимся очевидным примером: если в некотором предложении перед лексически омонимичной словоформой типа ***механику*** <***кибернетику, технику, физику, математику...***> стоит предлог ***в***, то она должна однозначно интерпретироваться как название области деятельности: МЕХАНИКА <КИБЕРНЕТИКА, ТЕХНИКА...> Если же перед такой словоформой стоит предлог ***к*** (***к механику*** <***кибернетику, технику, физику, математику...***>), то ее следует интерпретировать как название специалиста: МЕХАНИК <КИБЕРНЕТИК, ТЕХНИК...>

Основное назначение блока ПредсинтА нашего ЛП состоит именно в разрешении лексической и грамматической омонимии словоформ по линейному контексту.

Кроме того, в некоторых случаях ПредсинтА решает другую, более сложную задачу: он устанавливает конкретные синтаксические связи между теми словами анализируемого предложения, лексические и синтаксические свойства и относительное линейное расположение которых позволяют однозначно сделать вывод о наличии между ними таких связей. Например, для того чтобы установить подчинение частицы НЕ по ограничительному СинтО непосредственно следующему за ней глаголу в личной форме или инфинитиве (*не пишет, не пишите, не писать* и т. д.), не нужно исследовать сколько-нибудь широкий контекст подобных словосочетаний: это можно сделать сразу же после того, как мы обнаружили, что в предложении такая последовательность словоформ присутствует.

Таким образом, в результате применения правил ПредсинтА комбинированная МорфС предложения перерабатывается в некоторый промежуточный объект, в котором, по сравнению с исходной структурой, существенно сокращена лексико-грамматическая омонимия и проведены некоторые синтаксические связи: это своего рода эмбрион будущего дерева СинтС предложения. Важно отметить, что идея установления "сильных", высоковероятных синтаксических связей по ближайшему линейному контексту, первоначально используемая только на этапе ПредсинтА, оказалась весьма плодотворной и в конечном итоге привела к созданию алгоритма установления высоковероятных связей, который успешно эксплуатируется на этапе собственно синтаксического анализа (подробнее об этом см. в разд. 4.8.2).

В заключение этого раздела следует сказать несколько слов еще об одной задаче, решаемой на этапе ПредсинтА: правила этого этапа могут приписывать некоторым словоформам предложения вспомогательные характеристики, используемые на последующих этапах обработки предложения. В частности, с помощью таких правил приписывается особая помета "зеро" тем словам, которые могут взять на себя роль вершины дерева СинтС в случае отсутствия в предложении глагола-связки БЫТЬ. Например, в предложениях

(3) *Задача уже решена,*

(4) *Кто начальник коммерческого отдела?,*

(5) *Газовая хроматография - широко используемый метод мониторинга окружающей среды*

помета "зеро" приписывается, соответственно, словам ***решена, начальник*** и ***метод.***

4.5.3. Образцы правил предсинтаксического анализа

Ниже приводятся примеры правил ПредсинтА на формальном языке - по одному на общие, трафаретные и словарные правила. Каждое правило сопровождается содержательным комментарием.

Всего в настоящее время в блоке предсинтаксического анализа используется, не считая словарных, около 40 правил.

4.5.3.1. Общее правило предсинтаксического анализа

Формальная запись правила

```
REG:PRESYNT.21    СТИРАНИЕ АДЪЕКТИВНОГО ОМОНИМА КАЧЕСТВЕН-
CHECK             НОГО НАРЕЧИЯ
1.1 =(X,КР)&НОМ(X,ADV)
N:Ø1       РЯДОМ С Х-ОМ  ЕСТЬ ПРИЛАГАТЕЛЬНОЕ ИЛИ НЕСВЯЗОЧ-
CHECK      НЫЙ ГЛАГОЛ
1.1 M-EQUN(X,Z,Ø,A,V)&^LEXR(Z,БЫТЬ,ОКАЗЫВАТЬСЯ,СТАНОВИТЬСЯ)
DO
1 STEROM:X
N:Ø2       ВБЛИЗИ ОТ Х-А ЕСТЬ ПРИЛАГАТЕЛЬНОЕ ИЛИ НЕСВЯЗОЧНЫЙ
CHECK      ГЛАГОЛ
1.1 M-EQUN(X,Z,3,A,V)&^LEXR(Z,БЫТЬ,ОКАЗЫВАТЬСЯ,СТАНОВИТЬСЯ)
1.2 ININT(X,Z,И1)/ININT(Z,X,И1)/ININT(X,Z,ИЛИ)/
     ININT(Z,X,ИЛИ)
DO
1 STEROM:X
```

Комментарии

Данное правило предназначено для разрешения лексико-грамматической омонимии словоформ ***плохо, хорошо, интересно*** и т.п., которые, вообще говоря, могут интерпретироваться либо как наречия (в предложениях типа ***Он выступил хорошо*** <***интересно***>), либо как краткие формы прилагательных среднего рода (в предложениях типа ***Это (было) хорошо*** <***интересно***>).

Данное правило относится к альтернативному (обобщенному) типу. Общее условие 1.1 ориентирует правило на адъективную словоформу и сообщает, что рассматриваемая единица должна стоять в краткой форме и иметь омоним среди наречий.

В зоне проверки первой альтернативы правила разбирается ситуация, когда непосредственно слева или справа от рассматриваемой словоформы (X) стоит несвязочная глагольная лексема Z (т.е. лексема, отличная от слов БЫТЬ, ОКАЗЫВАТЬСЯ и СТАНОВИТЬСЯ).

В зоне проверки второй альтернативы рассматривается ситуация, когда несвязочная глагольная лексема или прилагательное Z стоит близко от рассматриваемой словоформы X (на расстоянии не более трех слов), но не рядом с ней. В этом случае дополнительно требуется (условие 1.2), чтобы в интервале между X и Z (или Z и X) находились сочинительные союзы И или ИЛИ.

Зоны действий в обеих альтернативах правила совпадают: содержащаяся в них инструкция стирает адъективный (т.е. именно тот, на который ориентировано правило) омоним X-а; тем самым данная омонимия решается в пользу наречия.

4.5.3.2. Трафаретное правило предсинтаксического анализа

Формальная запись правила

```
REG:PRESYNT.16        ПРИПИСЫВАНИЕ ХАРАКТЕРИСТИКИ ЗЕРО ЛОКА-
LOC:ALFA              ТИВНОМУ НАРЕЧИЮ ИЛИ ПРЕДЛОГУ; X=PR,ЛО-
ALFA:PR/ADV           КАТ/ADV,ЛОКАТ
N:Ø1
CHECK
1.1 =(X,ALFA,ЛОКАТ)&M-EQU(X,Z,6,S,ИМ)
1.2 ^ININT(X,Z,ЛИЧ)&^ININT(Z,X,ЛИЧ)
DO
1 DOBUZHAR:X(ЗЕРО)
```

Комментарии

Назначение данного правила - приписать уже упоминавшуюся характеристику "зеро" локативным наречиям типа ГДЕ, ТАМ, ДОМА, ВЕЗДЕ, (ПО)ВСЮДУ, СЛЕВА или локативным предлогам типа В2, НА2, НАД, ПОД2 в случае, если в рассматриваемом предложении отсутствует глагол-связка (таковы, например, конструкции типа ***Где ваш учитель?, Игорь сейчас дома, На столе белая скатерть***). Ссылка на это правило дается в словарных статьях локативных наречий и предлогов.

Правило носит безальтернативный характер.

В зоне проверки правила приводятся два условия: в усло-

вии 1:1 требуется, чтобы справа или слева на расстоянии не более шести словоформ от рассматриваемой локативной словоформы X имелось существительное в именительном падеже Z; условие 1.2 запрещает появление между X и Z глагола в личной форме.

Инструкция в зоне действий правила приписывает словоформе X помету "зеро", с помощью которой в дальнейшем блок синтаксического анализа сможет установить в предложениях рассматриваемого типа предикативное СинтО.

4.5.3.3. Словарное правило предсинтаксического анализа

Формальная запись правила

```
REG PRESYNT ØØ          СТИРАНИЕ ОМОНИМА В2 В ОТСУТСТВИЕ
N.Ø1                    ПРЕДЛОЖНОГО И МЕСТНОГО ПАДЕЖЕЙ
CHECK
1 1 ^R-EQUN(X,*,1Ø,ПР,МЕСТН)
DO
1 STEROM.X
```

Комментарии

Данное простое правило, помещенное непосредственно в словарной статье предлога В2 (который управляет предложным или местным падежом, как в конструкциях типа ***в поле*** [= пр. пад.] или ***в лесу*** [= местн. пад.]), стирает рассматриваемый омоним в том случае, если справа от этого предлога на разумном расстоянии (до 10 слов) нет ни одной именной словоформы в предложном или местном падежах (т.е. нет кандидата на заполнение валентности данного предлога).

Аналогичные правила помещены и в словарных статьях других предлогов; в частности, в статье предлога В1 (который управляет винительным падежом, как в конструкциях типа ***в дом***). Тем самым в значительном большинстве ситуаций предложная омонимия снимается на этапе ПредсинтА.

4.6. Синтаксический анализ в собственном смысле

В данном разделе мы рассмотрим правила синтаксического анализа в собственном смысле. Как уже отмечалось, этот этап ЛП является ключевым для описываемой системы, и используемые в нем правила представляют принципиальный лингвистический интерес.

Напомним, что в соответствии с блок-схемой синтаксического анализа, приведенной в разд. 4.4, входом блока СинтА в собственном смысле является модифицированная МорфС предложения - результат работы только что описанного блока ПредсинтА.

4.6.1. Типы синтаксических правил. Понятие синтагмы

Посредством правил СинтА на этой структуре формируется набор гипотез о возможных синтаксических связях между составляющими ее элементами. Основным критерием, лежащим в основе правил, формирующих синтаксические гипотезы (эти правила называются **синтагмами**), является критерий максимальной согласованности связанных лексем по всем типам приписанной им комбинаторной информации - как морфологической, так и лексикографической. Информация первого типа извлекается из МорфС, в которой она была выработана в процессе морфологического анализа; лексикографическая информация (включая синтаксическую, семантическую, сочетаемостную) извлекается из соответствующих статей комбинаторного словаря.

После того, как формирование множества синтаксических гипотез закончено, алгоритм синтаксического анализа приступает к его оптимизации, устраняя из этого множества ложные гипотезы на основе некоторых универсальных и локальных требований к правильной синтаксической структуре предложения.

Помимо синтагм - основного типа правил блока СинтА, в этом блоке обращается ограниченное количество (не более 10) довольно общих правил предпочтения, назначение которых состоит в установлении некоторых приоритетов для синтаксических гипотез, сформированных на пространстве обрабатываемого предложения. Скажем, если для словоформ X и Y анализируемой фразы сформированы гипотезы X $\xrightarrow{\text{1-компл}}$ Y и X $\xrightarrow{\text{атриб}}$ Y, из которых по условиям древесности должна быть оставлена только одна, стирается вторая гипотеза: актантная (в данном случае - 1-я комплетивная) связь, при прочих равных условиях, предпочитается атрибутивной связи как более сильная.

Поскольку правила предпочтения носят по преимуществу технический характер, мы считаем возможным не приводить их непосредственно в тексте настоящей книги, а ограничиться характеристикой синтагм.

6 Заказ № 1538

Всего в настоящее время в блоке собственно синтаксического анализа используется 30 общих и 284 трафаретных синтагмы; в словарных статьях комбинаторного словаря имеется еще около 150 синтагм.

Общие синтагмы, составляющие менее одной десятой части всего массива, касаются больших классов слов и используются при СинтА любой фразы, поскольку вероятность применения такой синтагмы весьма велика.

Трафаретные синтагмы - самый многочисленный класс синтаксических правил - касаются замкнутых и не слишком больших групп слов; что касается словарных синтагм, то они затрагивают так называемые грамматические слова, характеризующиеся особой "синтаксической чувствительностью" (это вспомогательные и модальные глаголы, местоимения разных разрядов, союзы, предлоги, частицы, а также полнозначные лексемы, обладающие какими-либо идиосинкратичными синтаксическими свойствами). Трафаретные и словарные синтагмы активируются лексическим составом обрабатываемого предложения.

Такая организация массива синтагм оправдана содержательно и имеет большой оптимизационный смысл, так как позволяет привлекать для анализа текущего предложения лишь те из синтагм, которые действительно имеют реальный шанс примениться. Это дает возможность при машинной реализации обойтись значительно меньшими компьютерными ресурсами и добиться существенного сокращения времени обработки предложения (по сравнению с ситуацией, когда массив правил никак не эшелонирован).

4.6.2. Образцы синтагм

Ниже приводятся примеры синтагм, записанных на формальном языке - по одной на каждый из трех названных типов - общие, трафаретные и словарные синтагмы. Каждая синтагма сопровождается подробным содержательным комментарием.

4.6.2.1. Общая синтагма

Формальная запись синтагмы

```
REG:АТРИБ 01
N:Ø1
CHECK
1.1 =(X,S)&R-EQUN(X,Y,1Ø,S,A,NUM)&=(Y,РОД)&#(Y,ХАРАКТРОД)&
     ^LEXR(Y,КОТОРЫЙ)&^ININT(X,Y,ЛИЧ,ДЕЕПР,ИНФ,ПОДЧ)
```

```
: 1 #(Y,A,NUM)/DOM(Y,*,ЭЛЕКТИВ)
:0
: SVUZOT (X,Y,АТРИБ)
```

Комментарии

Данная синтагма предназначена для установления гипотетического атрибутивного СинтО в конструкциях типа *библиотека* [X] *нашего района* [Y], *костюм* [X] *деда* [Y] и т.д. (здесь и далее в примерах буквами X и Y обозначаются те слова предложения, которые соответствуют одноименным лексическим переменным в синтагмах).

Зона проверки синтагмы содержит два условия.

Условие 1.1, относящееся к первой группе (т.е. к группе необходимых линейных условий; см. разд. 2.3), содержит следующие требования:

1. Словоформа X – потенциальный хозяин атрибутивного СинтО - должна быть существительным.

2. Справа от X на расстоянии не более десяти слов должно стоять существительное, прилагательное или числительное Y в родительном падеже - это потенциальный зависимый член атрибутивного СинтО.

3. Y не должен иметь синтаксического признака "характрод", т.е. он не должен быть словом типа *величина, размер, рост, толщина, цвет, форма,* которые в роли зависимых атрибутивного СинтО имеют ряд специфических свойств и поэтому обрабатываются отдельной синтагмой (см. разд. 4.6.2.2).

4. Y не должен быть словом КОТОРЫЙ. В рассматриваемой функции это слово тоже имеет синтаксическую специфику и тоже обрабатывается отдельным правилом (помещаемым непосредственно в его словарной статье).

5. В интервале между X-ом и Y-ом не должно быть личных форм глагола, деепричастий, инфинитивов, подчинительных союзов. Авторы синтаксической модели исходят из того, что такие элементы практически не могут (по крайней мере в текстах нейтрально-делового стиля) вклиниваться между элементами атрибутивной конструкции, ср. невозможность **автомашина, которую мне показали, директора* (при правильности *автомашина директора, которую мне показали*).

Условие 3.1, относящееся к группе необходимых древесных условий, гласит следующее: если Y, т.е. зависимый член

атрибутивной конструкции, является прилагательным или числительным, то он должен быть вершиной элективного отношения; ср. *дом первого из них <пяти из них>*

Зона действий данной атрибутивной синтагмы содержит единственную инструкцию (таковы зоны действий всех синтагм), смысл которой состоит в следующем: связать слово X в качестве синтаксического хозяина со словом Y в качестве зависимого элемента гипотетическим атрибутивным СинтО.

4.6.2.2. Трафаретная синтагма

Формальная запись синтагмы

```
REG:АТРИБ.1О
LOC:ALFA,R,R1,R2
ALFA·ЛИЧ/ДЕЕПР/ИНФ/ПОДЧ
R:ОПРЕД/АТРИБ/КВАЗИАГЕНТ/1-КОМПЛ/КОМПЛ-АППОЗ/КОЛИЧЕСТ
R1·ОПРЕД/АТРИБ
R2:ОПРЕД/КОЛИЧЕСТ
N:Ø1
CHECK
1.1 =(X,РОД)
3.1 L-EQU(X,Y,5,S)&^ININT(Y,X,ALFA)/R-EQU(X,Y,5,S)&^ININT
     (X,Y,ALFA)&L-DOM(Y,W,1Ø,ОПРЕД)&ORD(W,X)
3.2 DOM(X,Z,R)
3.3 #(X,ВРЕМ)/DOM(X,Z,R1)
3.4 #(X,'ВЕЩЕСТВО','СВОЙСТВО')/DOM(X,Z,ОПРЕД)
3.5 #(X,ПАРАМ)/DOM(X,Z,R2)
DO
1 SVUZOT·(Y,X,АТРИБ)
```

Комментарии

Данная синтагма применяется для анализа атрибутивной конструкции с существительным X, имеющим синтаксический признак "характрод", в качестве зависимого элемента, например: ***предложения*** [Y] ***такого*** [Z] ***вида*** [X], ***ток*** [Y] ***низкой*** [Z] ***частоты*** [X], ***языки*** [Y] ***высокого*** [Z] ***уровня*** [X], ***признаки*** [Y] ***группы*** [X] ***А*** [Z], ***старинное*** [W] ***времен*** [X] ***Ивана*** [Z] ***третьего оружие*** [Y].

Ссылки на данную синтагму (в виде записи TRAF:АТРИБ.10) даются в комбинаторном словаре, в словарных статьях тех существительных, которые обладают этим признаком.

Зона проверки этой синтагмы содержит одно условие первой группы (группы необходимых линейных условий) и пять условий третьей группы (группы необходимых древесных условий).

Условие 1.1 требует, чтобы слово X - потенциальный зависимый член атрибутивного СинтО - стояло в родительном падеже.

Условие 3.1 содержит следующие требования.

1. Слева или справа от X-а на расстоянии не более 5 слов должно быть существительное Y - потенциальный хозяин атрибутивного СинтО.

2. Между X-ом и Y-ом (или Y-ом и X-ом - в зависимости от того, как именно в обрабатываемом предложении располагаются кандидаты в члены СинтО) не должно быть личных форм глагола, деепричастий, инфинитивов, подчинительных союзов. Причины здесь те же, что и в предыдущей общей атрибутивной синтагме).

3. В случае, если Y стоит справа от своего гипотетического слуги X, то при Y-е должно быть препозитивное определение W, предшествующее синтаксической группе X-а (ср. ***новенькие*** [W], ***последних марок*** [X] ***машины*** [Y], при сомнительности ?***последних марок машины, экспонированные на выставке, привлекли внимание западноевропейских фирм***).

В условии 3.2 говорится, что у X-а должно быть определительное, атрибутивное, квазиагентивное, 1-е комплетивное, комплетивно-аппозитивное или количественное зависимое Z (***человек*** [Y] ***высокого*** [Z] ***роста*** [X], ***ружье*** [Y] ***образца*** [X] ***1930 года*** [Z], ***богатые старики*** [Y] ***его*** [Z] ***круга*** [X], ***птицы*** [Y] ***отряда*** [X] ***воробьиных*** [Z] и т.п.).

Условие 3.3 требует, чтобы в случае, когда X - кандидат в зависимые элементы конструкции - является существительным с синтаксическим признаком "врем" (т.е. временным существительным), любое его зависимое Z было только определительным или атрибутивным (***мебель*** [Y] ***петровской*** [Z] ***эпохи*** [X], ***мебель*** [Y] ***эпохи*** [X] ***Петра*** [Z]).

Условие 3.4 устанавливает, что в тех случаях, когда X обладает одним из семантических признаков (дескрипторов) 'вещество' или 'свойство', любое его зависимое Z может быть только определительным (ср. ***браслет*** [Y] ***чистого*** [Z] ***золота*** [X] при невозможности **браслет*** [Y] ***золота*** [X] ***восемьдесят шестой пробы*** [Z]).

Наконец, условие 3.5 утверждает, что если X – параметрическое существительное (типа ***высота, глубина, размер, рост*** и т.п.), то зависимое Z при нем может быть определительным или количественным (***перчатки*** [Y] ***большого*** [Z] ***размера*** [X], ***костюмы*** [Y] ***двух*** [Z] ***ростов*** [X]).

Зона действий данной трафаретной атрибутивной синтагмы содержит инструкцию, в некотором роде обратную той, какую мы имели в общей атрибутивной синтагме: связать слово Y в качестве синтаксического хозяина со словом X в качестве зависимого элемента гипотетическим атрибутивным СинтО.

Такая обратная ориентация синтагмы (относительно зависимого, а не главного члена конструкции) обусловлена тем фактом, что именно слова - кандидаты в зависимые члены - здесь легко идентифицировать в качестве принадлежащих к определенному синтаксическому типу; что же касается главного члена синтагмы, то он реализуется гораздо более широким классом слов.

4.6.2.3. Словарная синтагма

Формальная запись синтагмы

```
REG ЭЛЕКТИВ ОО
LOC ALFA,R
ALFA ПРЕВ/ПОРЯДК/ЭЛЕКТ/КОЛИЧЕСТ
R ОПРЕД/КОЛИЧЕСТ
N Ø1
CHECK
1 1 L-EQU(X,Y,Ø,ALFA)/R-EQU(X,Y,*,ALFA)
2 1 I-EQU(X,*,Y,ЛИЧ)
3 1 R-DOM-EQUN(X,Z,*,ПРЕДЛ,S,МН,КЛАС)
3 2 =(Z,МЕСТ)/=(Y,МН)/COGEN(Y,Z)
3 3 ORD(Y,X)/¯DEP-EQU(Y,*,R,S)&¯R-DOM(Y,*,2,1-КОМПЛ)
DO
1 SVUZOT (Y,X,ЭЛЕКТИВ)
```

Комментарии

Эта синтагма целиком записывается в словарной статье предлога ИЗ, который выступает в качестве зависимого члена X элективной конструкции. Она применяется для анализа конструкций типа ***интереснейшая*** [Y] ***из*** [X] ***книг*** [Z], ***первое*** [Y] ***из*** [X] ***утверждений*** [Z], ***миллион*** [Y] ***из*** [X] ***книг*** [Z], ***пять***

[Y] *из* [X] *этих книг* [Z], *последний* [Y] *из* [X] *людей* [Z], *некоторые* [Y] *из* [X] *свойств* [Z], *какой-то* ‹*какой-нибудь*› [Y] *из* [X] *аккумуляторов* [Z], *кто-то* [Y] *из* [X] *вас* [Z]

Зона проверки синтагмы состоит из пяти условий: одного необходимого линейного условия, одного невозможного линейного условия и трех необходимых древесных условий.

Условие 1 1 требует, чтобы непосредственнно слева от X-а или на некотором расстоянии справа от него имелась словоформа Y – прилагательное в превосходной степени, либо порядковое прилагательное (*первый, второй, третий,...*), либо прилагательное или существительное, обладающее синтаксическим признаком "элект" (*последний, многие, некоторый, какой,...; тысяча, миллион,...*), либо, наконец, числительное.

Условие 2. 1 запрещает ситуацию, при которой Y находится справа от X-а и при этом между X-ом и Y-ом имеется глагол в личной форме. Дело в том, что появление в промежутке между X-ом и Y-ом личного глагола влечет за собой непроективность конструкции (перекрытие вершины, ср.

электив
↓ ┌──────────────────────┐
Из [X] *них слева находились* предик→ *пять* [Y]) Непроективные конструкции описываются особыми синтагмами.

Условие 3. 1 гласит, что справа от X-а должно находиться подчиненное ему (по предложному СинтО) существительное Z, стоящее во множественном числе или имеющее синтаксический признак "множественности" (ср. КЛАСС, ГРУППА, СЕМЬЯ, ПРАВИТЕЛЬСТВО и т. п. : *первый из их класса*).

В условии 3. 2 требуется, чтобы либо Z являлось местоименным существительным (*лучший* [Y] *из* [X] *них* [Z]), либо Y стояло в форме множественного числа (*лучшие* [Y] *из* [X] *сотрудников* [Z]), либо между Y-ом и Z-ом выполнялось согласование по роду (*лучший* [Y] *из* [X] *учеников* [Z], *лучшая* [Y] *из* [X] *ее подруг* [Z], *лучшее* [Y] *из* [X] *его творений* [Z]).

Наконец, условие 3. 3 утверждает, что либо Y предшествует X-у, либо Y является определительным или количественным зависимым какого-нибудь существительного и при этом справа от Y-а не стоит слово, зависящее от него по первому комплетивному СинтО (тем самым отвергается как неграмматичная конструкция типа **тысяча книг из них*).

Зона действий синтагмы содержит инструкцию, в соответствии с которой устанавливается гипотетическое элективное

СинтО между словом Y в качестве хозяина и предлогом ИЗ [=X] в качестве зависимого члена.

4.7. Правила синтаксического синтеза

До сих пор мы рассматривали правила синтаксического компонента ЛП, работающие в направлении анализа, т.е. предназначенные для построения СинтС предложения по его МорфС.

Ниже будут коротко рассмотрены правила синтаксического синтеза, с помощью которых решается обратная задача: построение МорфС предложения по его СинтС. Если вернуться к сказанному в разделе 4.2 и вспомнить о тех различиях, которые существуют между СинтС и МорфС предложения, то будет нетрудно понять, какие основные проблемы приходятся на долю этапа СинтСз: это 1) морфологизация СинтС (т.е. приписывание ее узлам тех морфологических характеристик, которых недостает, чтобы выбрать из парадигмы лексемы, соответствующей данному узлу, нужную словоформу); 2) линеаризация дерева СинтС и 3) расстановка знаков препинания в выходной МорфС.

Необходимо отметить, что в основной задаче нашего ЛП - системе общения с информационной базой на русском языке - правила СинтСз не используются. В дальнейшем, когда мы будем в той или иной форме генерировать ответ системы на русском языке, эти правила найдут применение. В настоящее же время они активно участвуют в системе машинного перевода с английского языка на русский.

Следуя уже использованной дважды методике, мы приведем три образца правил СинтСз - по одному на каждый из трех типов правил. Как и в предыдущих разделах, правила вначале даются в формальной записи, а затем снабжаются содержательными комментариями.

4.7.1. Общее правило

Формальная запись правила

```
REG:SYNTHES2.22
CHECK
1.1 DOM-EQUN(X,Z,ПРЕДИК,S,A,NUM)
N:Ø1
CHECK
1.1 ^#(Z,МН,ВЕРШСОЧ,NUM)&#(Z,ОДИН,ДВАДОД)/DOM-NEQUN(Z,*,
```

```
     КОЛИЧЕСТ,ОДИН,ДВАДОД,ЦИФ-1)
Ↄ0
: DOBUZHAR:X(МН,З-Л)
N:Ø2
CHECK
Ↄ0
1 PERUZHAR:Z(RGNR)-X
2 ZAMUZHAR:X(ЕД,З-Л)
```

Комментарии

Данное правило, используемое для синтеза характеристик рода, числа и лица сказуемого, принадлежит к типу обобщенных правил и включает два альтернативных подправила.

Общее условие правила требует, чтобы подлежащее Z – зависимый член предикативной конструкции, подчиненный сказуемому X, - было существительным, прилагательным или числительным.

Зоны условий, а также зоны действий в двух альтернативных подправилах различны.

Единственное условие, содержащееся в первом подправиле, требует, чтобы подлежащее Z являлось либо 1) числительным (за исключением числительного ОДИН или составного числительного, оканчивающегося на ОДИН, например, ***двадцать один***); либо 2) количественной группой с таким числительным (***два завода***); либо 3) именем во множественном числе (***заводы***); либо 4) вершиной сочинительной цепочки (***завод и фабрика***).

Если это условие выполнено, то с помощью инструкции, содержащейся в зоне действий данного подправила, сказуемому X приписываются характеристики множественного числа и третьего лица. Тем самым фрагмент структуры типа $\text{ЗАВОД}_{\text{мн}} \xleftarrow{\text{предик}} \text{РАБОТАТЬ}$ превращается во фрагмент структуры типа $\text{ЗАВОД}_{\text{мн}} \xleftarrow{\text{предик}} \text{РАБОТАТЬ}_{\text{мн,3-л}}$ (который в дальнейшем, в ходе морфологического синтеза, превратится в словосочетание ***заводы работают***).

В случае, если общее условие выполняется, а условия первого подправила - нет, то правило переходит к исполнению инструкций, предусмотренных в зоне действий второго подправила (поскольку зона условий данного подправила пуста). Эти инструкции, во-первых, передают характеристику рода от под-

лежащего Z сказуемому X и, во-вторых, приписывают сказуемому X характеристики единственного числа и третьего лица. Этим подправилом, например, фрагмент конструкции типа ЗАВОД$_{муж,ед}$ $\xleftarrow{предик}$ РАБОТАТЬ превращается во фрагмент ЗАВОД$_{муж,ед}$ $\xleftarrow{предик}$ РАБОТАТЬ$_{муж,ед,3-л}$ (и в дальнейшем в словосочетание *завод работает*).

Нужно отметить, что хотя приведенное правило СинтСз SYNTHES2.22 описывает важный и сложный фрагмент русского синтеза, оно имеет достаточно простой и компактный вид - благодаря тому, что оно (как и другие общие правила) игнорирует нестандартные случаи, например, согласование сказуемого с подлежащим в контексте отрицания (типа *газет* [жен,мн] *не было* [ср,ед]); согласование сказуемого с личными местоимениями 1-го и 2-го лица (Я, МЫ, ТЫ, ВЫ). С этими предикативными конструкциями мы поступаем следующим образом: вначале они обрабатываются общим правилом SYNTHES2 22, а затем словарные правила СинтСз (которые работают после общих правил) вносят необходимые коррективы. В разд. 4.7.3 в качестве примера словарного правила приводится одно из этих корректирующих правил.

4.7.2. Трафаретное правило

Формальная запись правила

```
REG:SYNTHES1 18
CHECK
1.1 DOM(X,Z,ПРИСВЯЗ)/DOM(X,Z,ПАСС-АНАЛ)
1.2 #(Z,ИМ,РОД,ДАТ,ВИН,ТВОР,ПР,МЕСТН)
N:Ø1
CHECK
1 1 ^#(X,ПРОШ,ИНФ,ДЕЕПР)/=(X,НЕПРОШ,ЛИЧ,СОВ)
1.2 ^#(Z,S,A,NID)&#(Z,КАЧЕСТВ)&^DOM(Z,*,КОЛИЧЕСТ)/=(Z,ПРИЧ)
     &#(Z,СТРАД)
DO
1 DOBUZHAR·Z(ТВОР)
N:Ø2
CHECK
1.1 =(Z,A,КАЧЕСТВ)&#(Z,СРАВ)/=(Z,ПРИЧ,СТРАД)
2.1 DOM(Z,*,ОПРЕД)
DO
1 DOBUZHAR Z(КР)
```

```
N:Ø3
CHECK
DO
1 DOBUZHAR Z(ИМ)
```

Комментарии

Данное трафаретное правило используется для синтеза падежа и краткой формы в присвязочной конструкции. Ссылка на это правило дается в словарных статьях связочных глаголов типа БЫТЬ, ОКАЗЫВАТЬСЯ и др.

Правило включает в себя три альтернативных подправила со следующим общим условием: у связочного глагола X имеется присвязочное или пассивно-аналитическое зависимое Z, не имеющее падежной характеристики.

Подправило 1 включает два условия:

1. Словоформа X имеет одну из следующих характеристик: прошедшее время; инфинитив; деепричастие; непрошедшее время совершенного вида.

2. Зависимое Z – это либо прилагательное (не имеющее признака "качеств"), либо существительное, у которого нет количественных зависимых, либо причастие (не страдательного залога).

Инструкцией из зоны действий данного подправила именной части сказуемого - Z-у - приписывается характеристика творительного падежа. Например, фрагмент конструкции типа ОН $\xleftarrow{\text{предик}}$ БЫТЬ$_{\text{прош}}$ $\xrightarrow{\text{присвяз}}$ ИНЖЕНЕР заменяется фрагментом ОН $\xleftarrow{\text{предик}}$ БЫТЬ$_{\text{прош}}$ $\xrightarrow{\text{присвяз}}$ ИНЖЕНЕР$_{\text{твор}}$ (что в ходе морфологического синтеза преобразуется в ***Он был инженером***).

Подправило 2 содержит единственное условие, требующее, чтобы Z было качественным прилагательным (не в сравнительной степени) или страдательным причастием.

Инструкция из зоны действий приписывает Z-у характеристику "краткость". Например, фрагмент БЫТЬ$_{\text{прош}}$ $\xrightarrow{\text{пасс-анал}}$ ПРИНЯТЬ$_{\text{прич,страд}}$ превращается во фрагмент структуры БЫТЬ$_{\text{прош}}$ $\xrightarrow{\text{пасс-анал}}$ ПРИНЯТЬ$_{\text{прич,страд,кр}}$ (и в дальнейшем в словосочетание ***был принят***).

Во всех остальных случаях (т.е. при невыполнении условий подправил 1 и 2) работает "дежурное" подправило 3, имеющее пустую зону проверки, инструкция из зоны действий которого

приписывает Z-у характеристику именительного падежа. Например, фрагмент структуры ДЕТАЛЬ - СТАЛЬНОЙ$_{жен}$ (предик) превращается в ДЕТАЛЬ - СТАЛЬНОЙ$_{жен,им}$ (предик), что впоследствии преобразуется в предложение *Деталь - стальная.*

4.7.3. Словарное правило

Формальная запись правила

```
REG SYNTHES2 ØØ
N Ø1
CHECK
1 1 DEP(X,Z,ПРЕДИК)
DO
1 ZAMUZHAR Z(1-Л)
```

Комментарии

Данное правило, помещенное в словарной статье лексемы МЫ, корректирует характеристику лица сказуемого при подлежащем МЫ, выработанную ранее общим правилом согласования подлежащего со сказуемым. Как уже говорилось, общие правила имеют простой и компактный вид, но не во всех случаях обеспечивают нужный результат и поэтому могут иногда требовать дальнейшей корректировки трафаретными или словарными правилами.

Единственное условие правила требует, чтобы слово X - МЫ - подчинялось слову Z по предикативному СинтО.

В этом случае с помощью инструкции из зоны действий характеристика лица сказуемого Z, какова бы она ни была, заменяется характеристикой первого лица. Например, фрагмент типа МЫ $\xleftarrow{предик}$ ЧИТАТЬ$_{мн,3-л}$ с помощью данного словарного правила превращается во фрагмент МЫ $\xleftarrow{предик}$ ЧИТАТЬ$_{мн,1-л}$.

4.8. Алгоритмы синтаксического анализа

Алгоритм синтаксического анализа текстов на ЕЯ был подробно описан в работе [Цинман и др., 1986]. В ходе дальнейших исследований он был дополнен алгоритмом поиска высоковероятных синтаксических связей, который в настоящее время находится в экспериментальной стадии. Это своего рода приставка к основному алгоритму, задачей которой является повышение скорости синтаксического анализа. Таким образом,

ЛП в его нынешнем виде использует два алгоритма синтаксического анализа, определенным образом взаимодействующие друг с другом. Они описываются в разд. 4.8.1 и 4.8.2 соответственно.

4.8.1. Основной алгоритм

4.8.1.1. Алгоритмическая процедура работы с правилом преобразования

Мы исходим из того, что подавляющая часть лингвистической информации, необходимой для работы ЛП, задается в виде правил преобразования. В предшествующих разделах данной главы были описаны типы правил, используемых в системе синтаксического анализа предложения, и приведены образцы правил предсинтаксического и собственно синтаксического анализа и синтеза

Правила - сложные формальные объекты, состоящие, как было сказано выше, из двух основных зон - зоны условий и зоны действий.

Условия в правилах представляют собой логические выражения, в записи которых может быть использовано около 50 элементарных и свыше 100 составных предикатов (см. о них гл. 2). С помощью этих предикатов можно описать наличие (или отсутствие) у слова определенных характеристик, всевозможные согласования между выделенными словами фразы, необходимый (или невозможный) линейный или древесный контекст выделенных слов фразы и т д.

Зона действий в правилах - это перечень инструкций, последовательное исполнение которых осуществляет требуемое правилом преобразование рассматриваемого объекта. В подавляющем большинстве правил синтаксического анализа объектом преобразования является морфологическая структура. Инструкции позволяют изменить набор характеристик отдельных слов, стереть или сохранить те или иные омонимы словоформы, построить бинарное поддерево из двух элементов фразы, стереть некоторые из уже построенных гипотетических связей. В принципе возможно использование любых других инструкций из числа тех, которые названы в гл. 2, но в рамках алгоритма синтаксического анализа в них не возникает нужды.

В записи как предикатов, так и инструкций широко используются переменные различных типов.

Сложность правил и их широкое использование на всех неморфологических этапах работы системы приводит к тому, что принципиальным местом алгоритмического обеспечения нашей системы становится создание оптимальной алгоритмической процедуры, обрабатывающей правила преобразования. При построении основных узлов этой процедуры были приняты некоторые нетривиальные решения, позволившие резко сократить огромный перебор, который возникает при проверке истинности громоздких логических выражений с большим числом различных переменных.

4.8.1.2. Алгоритм предсинтаксического анализа

Этот этап анализа включен в систему из прагматических соображений. Предсинтаксические правила преобразования позволяют в ряде случаев по близкому линейному контексту стереть некоторые омонимы у слов анализируемой фразы, что приводит к существенному сокращению возможных синтаксических гипотез на этапе собственно синтаксического анализа и, следовательно, к ускорению процедуры построения СинтС обрабатываемой фразы. С алгоритмической точки зрения этот этап работы крайне прост и сводится к последовательному применению описанных в разд. 4.5 правил предсинтаксического анализа - как общих, так и трафаретных и словарных.

4.8.1.3. Алгоритм синтаксического анализа

Синтаксический анализ фразы является наиболее сложным этапом работы ЛП.

Напомним, что на вход этого этапа поступает морфологическая структура фразы, а на выходе должна быть получена ее синтаксическая структура - дерево зависимостей, в узлах которого стоят слова анализируемой фразы, а ветви помечены именами синтаксических отношений. Основным инструментом, предназначенным для построения СинтС, являются синтагмы, т.е. правила преобразования, которые при выполнении указанных в них условий позволяют связать два слова фразы некоторым СинтО.

Отметим несколько важных особенностей принятого в нашей системе описания синтаксиса входного языка в виде синтагм:

1) синтагма утверждает лишь возможность проведения СинтО между двумя словами, но не обязательность этой связи;

2) в большей части синтагм условия, описывающие кон-

текст, сформулированы в терминах наличия или отсутствия некоторых СинтО, связывающих слова рассматриваемой пары с другими словами фразы (древесный контекст);

3) множество синтагм невозможно упорядочить таким образом, чтобы в каждой синтагме древесный контекст задавался посредством СинтО, установленных ранее рассмотренными синтагмами.

Такое задание синтаксиса не представляет никакого регулярного механизма для построения дерева зависимостей. Синтагмы скорее предназначены для проверки правильности готовой СинтС. Поэтому в нашей системе принят метод построения СинтС, при котором вначале создается заведомо избыточный набор гипотетических связей, а затем ложные гипотезы отсеиваются с помощью дополнительных проверок и различных фильтров.

Как уже говорилось, синтагмы бинарны в том смысле, что каждая из них завершается единственной инструкцией, предлагающей связать посредством СинтО два и только два слова фразы, которые обозначаются в синтагмах через X и Y

Все условия синтагм подразделяются на четыре группы. Группы с номерами 1 и 3 состоят из необходимых, а с номерами 2 и 4 - из невозможных условий.

Условия групп 1 и 2 описывают свойства слов X и Y и линейный контекст этих слов во фразе (их относительный порядок, расстояние между ними, наличие или отсутствие между X и Y или слева или справа от них слов с определенным набором характеристик, знаков препинания и т. д.). Истинность или ложность этих условий может быть установлена прямым просмотром входной фразы.

Условия групп 3 и 4 описывают непосредственный древесный контекст слов X и Y, т.е. свойства слов или цепочек слов, синтаксическая связь которых со словами X и Y задается в явном виде. Начинать проверку этих условий можно лишь после того, как установлены все гипотетические связи между словами фразы.

Разделение условий синтагм на группы производится при написании синтагмы.

Как уже не раз отмечалось, весь корпус синтагм разбит на три части: общие, трафаретные и словарные синтагмы. Общие синтагмы участвуют в обработке каждой фразы. Из трафаретных

и словарных синтагм работают лишь те, на которые имеется ссылка в словарных статьях слов анализируемой фразы.

Блок 1: построение набора гипотетических СинтО

На первом этапе СинтА гипотетические СинтО между словами устанавливаются на основе проверки только тех условий синтагм, которые описывают линейный контекст (условия групп 1 и 2). Условия на древесный контекст на этом этапе не проверяются.

В результате возникает некоторый ориентированный граф гипотетических СинтО между словами фразы. Этот граф состоит из N узлов, где N - число слов анализируемой фразы. Сведения о графе удобно хранить в матрице размерности N × N При этом в клетке матрицы с координатами (p,q) содержатся сведения обо всех гипотетических СинтО, связывающих какой-либо омоним слова с номером p (синтаксический хозяин) с каким-либо омонимом слова с номером q (синтаксический слуга). Естественно, что число построенных таким образом гипотетических СинтО заметно превышает количество СинтО, необходимое для построения СинтС. СинтС фразы из N слов должна содержать N-1 СинтО. Реальное же число дуг (ветвей) графа обычно в 2-4 раза превосходит это число.

Основная сложность проверки условий синтаксических правил - широкое использование в этих условиях переменных разных типов. Процедура проверки условий сводится к поиску такого набора значений переменных, при котором условия становятся истинными. Подобный набор значений переменных для необходимых условий называется подтверждающим примером, а для невозможных условий - опровергающим примером. Перебор значений переменных осуществляется некоторым регулярным образом в предикатах, где эти переменные встретились впервые. Для всех гипотетических СинтО информация о данных переменных сохраняется. При этом про каждую переменную запоминается не только подобранное для нее значение, но и точное место в правиле, где данная переменная появилась впервые.

Блок 2: определение вершины СинтС

После построения графа гипотетических СинтО все усилия направляются на то, чтобы из графа выделить дерево - СинтС исходной фразы. Этот процесс происходит значительно быс-

трее, если сразу удается определить вершину дерева зависимостей. Поэтому после построения графа начинает работать блок, задача которого - выявление слов, являющихся кандидатами на роль вершины дерева.

Описание свойств слов, которые могут быть вершинами в СинтС фразы, задано в виде специального правила. Стандартный аппарат работы с правилом просматривает все омонимы всех слов фразы и помечает те из них, которые по своим характеристикам могут быть вершиной. Более того, с помощью приписывания некоторого веса кандидаты на роль вершины упорядочиваются в порядке убывания вероятности того, что это слово будет выбрано вершиной СинтС.

Затем осуществляется просмотр графа с целью выявить те его узлы, в которые не входит ни одна дуга. Таким узлам соответствуют слова фразы, ни один из омонимов которых не имеет ни одного синтаксического хозяина. Если таких слов больше одного, то граф несвязный и, стало быть, искомого дерева зависимостей получить не удается. Если имеется ровно одно такое слово, но ни один его омоним не помечен как кандидат на роль вершины, то и в этом случае нет надежды на построение правильной СинтС. Если же у этого слова есть омонимы с нужными пометами, то вершиной дерева должен быть один из этих омонимов. Поэтому в дальнейшем мы будем испытывать их в порядке убывания приписанного им веса.

Если, наконец, в каждый узел графа входит хотя бы одна дуга, то возникает необходимость организовать более обширный перебор, при котором просматриваются все кандидаты на роль вершины. Таким образом, определение вершины СинтС - первое место в алгоритме, где может возникнуть ветвление, связанное с рассмотрением различных вариантов синтаксического разбора фразы. Если выбранный вариант разбора не приводит к построению СинтС, происходит возврат в эту точку алгоритма. Поэтому все данные, находящиеся в оперативной памяти, которые при дальнейшем анализе фразы могут быть изменены, должны перед началом перебора кандидатов в вершины сохраняться на диске с тем, чтобы при необходимости их прежние значения можно было восстановить. Разбор одного варианта заключается в том, что очередной кандидат объявляется вершиной; это означает, что все омонимы данного слова, кроме рассматриваемого, уничтожаются (вместе с гипотетичес-

Заказ № 1538

кими СинтО, в которых они участвуют); стираются также дуги графа, входящие в выбранную вершину (если таковые были).

Блок 3: фильтры по окончательным и квазиокончательным СинтО

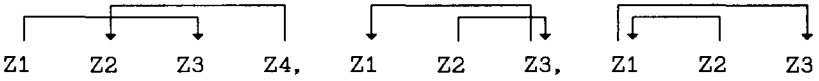
В этом блоке производится просмотр графа гипотетических СинтО. При этом совершаются следующие действия.

1. Если у некоторого слова (отличного от вершины) есть омонимы, не имеющие ни одного синтаксического хозяина, то эти омонимы стираются. Стираются также все гипотетические СинтО, в которых эти омонимы участвуют.

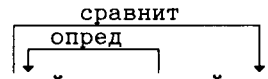
2. Если в графе имеется узел, куда входит только одна дуга, то она помечается как окончательная. В самом деле, эта дуга непременно должна принадлежать искомому дереву, иначе граф окажется несвязным.

3. Если в графе имеется узел, куда входит несколько дуг, но все они исходят из одного и того же узла графа, то эти дуги помечаются как квазиокончательные. Действительно, эти два узла в искомом дереве должны быть обязательно связаны одной из этих дуг (хотя неизвестно, какой). Это довольно распространенная ситуация, когда слово еще представлено несколькими омонимами, каждый из которых независимо участвовал в процессе построения гипотетических СинтО.

Появление в графе окончательного СинтО позволяет осуществить некоторую чистку графа. Прежде всего стираются все омонимы той пары слов, элементы которой связаны окончательной связью, кроме омонимов, образующих именно данную связь. Далее происходит обращение к довольно мощному фильтру, каковым является требование проективности СинтС. Известно, что абсолютное большинство правильно построенных фраз русского языка (а также ряда других европейских языков, в том числе английского и французского) имеют проективные СинтС. Это означает, что если r_1 и r_2 – проективные СинтО, то для них запрещена ситуация пересечения (независимо от направления стрелок). Кроме того, если r – проективное СинтО, соединяющее слова X и Y, то все слова, расположенные между X и Y, должны непосредственно или опосредованно подчиняться X и Y. Отсюда следует, что проективное СинтО не может огибать вершину фразы (ситуация обрамления стрелок). Иными словами, невозможны конфигурации вида



Некоторое (небольшое) число синтаксических конструкций ЕЯ непроективно. Таковы, например, русские сравнительные конструкции типа



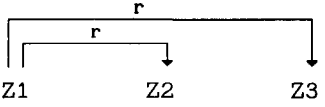
такой интересный, что все заслушались,

в которых имеет место непроективная ситуация обрамления стрелок.

В нашей системе допускается возникновение непроективных конструкций. Если известно, что порождаемые некоторой синтагмой СинтО могут нарушать проективность, в соответствующей синтагме должна быть сделана специальная помета.

Наконец, еще один фильтр: проверяется, принадлежит ли имя установленного окончательного СинтО к списку неповторимых СинтО.

СинтО r называется неповторимым, если у слова в СинтС не может быть двух синтаксических слуг, связанных с ним этим отношением, т. е. если невозможна ситуация



Почти все СинтО в русском языке обладают таким свойством. Тем самым в случае неповторимости окончательного СинтО можно у хозяина этого отношения стереть всех других слуг, связанных с ним тем же отношением.

Выявление окончательных СинтО является, таким образом, достаточно мощным фильтром. Поэтому в последней версии нашей системы была реализована возможность объявлять некоторые СинтО окончательными сразу после рассмотрения соответствующих синтагм.

Блок 4: проверка древесного контекста

Напомним, что гипотетические СинтО создавались на основе проверки условий синтагм, описывающих только линейный контекст (условия первой и второй групп). После получения графа гипотетических СинтО можно перейти к проверке древесного контекста (условия третьей и четвертой групп).

Проверка этих условий также представляет собой процедуру построения подтверждающих (опровергающих) примеров для групп необходимых (невозможных) древесных условий, причем для переменных, которые уже встречались в линейных условиях, берутся ранее найденные значения.

Отметим принципиальную трудность, возникающую при работе с синтагмами. Если проверка условий синтаксического правила производится на готовой СинтС, то наличие (отсутствие) подтверждающего примера означает истинность (ложность) группы необходимых условий, а отсутствие (наличие) опровергающего примера означает истинность (ложность) группы невозможных условий. Иначе обстоит дело на этапе СинтА, где вместо дерева приходится просматривать граф гипотетических СинтО. Лишь отсутствие подтверждающего (опровергающего) примера означает ложность (истинность) соответствующей группы необходимых (невозможных) условий. В то же время наличие подтверждающего примера для группы необходимых условий, вообще говоря, не означает их истинности, так же как и наличие опровергающего примера для группы невозможных условий не означает их ложности, потому что в процессе снятия омонимии и уничтожения избыточных гипотез контекст может уточниться и найденные примеры перестанут существовать.

Для учета этих обстоятельств некоторые предикаты, занятые подбором значений переменных, на этапе СинтА могут вырабатывать, помимо значений "истина" и "ложь", третье истинностное значение "квазиистина". Значение "квазиистина" вырабатывается предикатами в тех случаях, когда в качестве значения переменной подобран некоторый узел графа, связанный с данным узлом требуемым СинтО, но это СинтО не является окончательным. Логические операции над предикатами - это операции трехзначной логики. Поэтому могут возникнуть "квазиистинные" условия. Заметим, что могут быть окончательные гипотезы, в которых некоторые условия еще "квазиистинны". С другой стороны, могут существовать неокончательные гипотезы, у которых все условия истинны.

После проверки условий третьей и четвертой групп часть гипотез отсеивается.

Блок 5: правила предпочтения

Блоки 3 и 4 работают в цикле до тех пор, пока фильтры,

описанные в этих блоках, производят изменения в графе (стираются какие-либо омонимы, удаляются гипотезы). Этот цикл обычно состоит из одного - трех проходов, после чего происходит стабилизация графа.

Если граф, оставаясь связным, деревом не является, то происходит обращение к правилам предпочтения. Эти правила позволяют в отдельных случаях путем сравнения гипотетических синтаксических хозяев и слуг каждого слова оказать предпочтение одним гипотетическим связям, уничтожив другие.

Правила предпочтения носят вероятностный характер и включены в систему из практических соображений для ускорения процесса построения СинтС. Если обработка правил привела к уничтожению некоторых гипотез, то вновь происходит обращение к фильтрам блоков 3 и 4, поскольку всякое изменение в графе может повлечь за собой и другие изменения.

Блок 6: рассмотрение альтернативных деревьев

Этот блок вступает в действие, если после работы предыдущих блоков наступила стабилизация графа, но он по-прежнему не является деревом. Это второе место в алгоритме (после выбора вершины), где осуществляется перебор различных вариантов дальнейшего анализа предложения.

Результатом работы этого блока должно стать построение СинтС фразы. Организовано это построение в виде итеративной процедуры, каждый шаг которой дает новый импульс к дальнейшей редукции графа гипотетических СинтО.

Типичный шаг процедуры состоит в следующем. Отыскивается узел, в который входит более чем одна дуга. Из этих дуг одна и только одна попадет в искомое дерево. Дуги упорядочиваются: сначала просматриваются более близкие во фразе гипотетические синтаксические хозяева данного узла, потом более далекие. Когда некоторая дуга, входящая в рассматриваемый узел, выбрана, остальные дуги стираются.

Эти действия могут вызвать другие изменения в графе, поэтому вновь включаются в работу фильтры из блоков 3 и 4. Если вновь наступает стабилизация графа, то происходит переход к следующему шагу итеративной процедуры: в графе отыскивается узел, в который входит более одной дуги. При этом, переходя к очередному шагу, необходимо сохранять всю информацию, которая может меняться в процессе чистки графа.

Действительно, если рассмотрение очередной альтернативы привело граф к утрате связности, следует вернуться к предыдущему состоянию, выбрав следующую дугу, входящую в рассматриваемый узел, или, если таковых не осталось, вернуться в еще более раннюю точку рассмотрения альтернатив - к узлу, с которым производилась работа на предшествующем шаге.

Первая из СинтС, построенная алгоритмом СинтА, подается на следующие этапы работы системы. При этом в принципе сохраняется возможность вернуться к этапу СинтА за новой СинтС, если построенная СинтС оказалась неприемлемой на одном из последующих этапах работы системы. Впрочем, эта возможность осталась практически не реализованной, поскольку без использования интерактивного режима трудно уточнить понятие неприемлемости СинтО.

Изложенный алгоритм, по замыслу, должен обязательно приводить к построению по крайней мере одной правильной СинтС для обрабатываемой фразы. Поскольку, однако, это алгоритм фильтрового типа, требующий полного перебора всех возможных гипотез, он работает относительно долго. Поэтому, как уже говорилось, была сделана попытка дополнить его алгоритмом поиска высоковероятных связей, который работает существенно быстрее, хотя, быть может, и не обеспечивает в общем случае построения СинтС для обрабатываемого предложения.

4.8.2. Алгоритм установления высоковероятных связей

4.8.2.1. Общие сведения

В этом разделе описывается подход к синтаксическому анализу, основанный на понятии фрагмента - отрезка фразы с фиксированным синтаксическим разбором. При соединении слов соседних фрагментов синтаксическими связями возникают более крупные фрагменты; таким способом может быть получена СинтС всей фразы. Этот подход, объединяющий идеологию деревьев зависимостей и непосредственных составляющих, был предложен Г. С. Цейтиным и описан в статьях [Лейкина, Цейтин, 1975; Крупко, Цейтин, 1978]. В работе [Митюшин, 1985] охарактеризован класс структур, которые могут быть построены таким способом; в работе [Митюшин, 1988] предложен основанный на данном подходе алгоритм частичного анализа, устанавливающий наиболее вероятные синтаксические связи.

В настоящее время алгоритм частичного фрагментного ана-

лиза реализован как модуль в рамках общей системы СинтА. Характеристики его работы оказались достаточно хорошими: скорость анализа составляет около 10 слов в секунду процессорного времени, при этом строится в среднем 70-80 % всех синтаксических связей фразы. Точная оценка надежности не проводилась, но можно утверждать, что число ошибочных связей не превышает 1 % (надежность существенно зависит от качества используемых синтагм).

Модуль частичного фрагментного анализа присоединен как предварительный этап к основной системе русского СинтА, описанной в разд. 4.8.1. Такое сочетание дает заметный выигрыш в скорости, особенно на более длинных фразах - для них время анализа сокращается в 2-3 раза. Если некоторая связь, построенная на предварительном этапе, оказывается ошибочной, основная система проводит синтаксический анализ заново, не учитывая результаты предварительного этапа.

В дальнейшем эффективность алгоритма фрагментного анализа предполагается существенно повысить, в результате чего доля устанавливаемых связей дойдет до 100 %. Тогда в основную систему фильтрового анализа будут подаваться полные СинтС, рассматриваемые как рекомендуемые или гипотетические. Функция основной системы будет состоять в их детальной проверке с использованием полного корпуса синтагм, а также в автономном проведении анализа в аварийных ситуациях, т.е. в случаях, когда гипотетические структуры оказываются неправильными или вообще не строятся. Скорость работы подобного "тандема" может быть на порядок выше, чем у автономной системы фильтрового анализа.

При фрагментном анализе возникает естественная возможность приписывать фрагментам и полным структурам предпочтения или оценки, отражающие их синтаксическое качество. Этот подход применялся Г.С.Цейтиным и его последователями [Цейтин, 1975; Крупко, Цейтин, 1978; Железняков, Крупко, 1980] и активно развивается в настоящее время (см., напр.: [Кулагина, 1987, 1990; Tsujii et al., 1988]). Неединственность синтаксической структуры - весьма типичное явление, и предпочтения позволяют выбирать наиболее правдоподобные из возможных вариантов. Предпочтения могут быть полезны также на промежуточных стадиях анализа, направляя процесс построения структуры по наиболее перспективному пути.

Выделение класса высоковероятных связей в неявной форме уже означает присвоение этим связям более высокого приоритета по сравнению с остальными. В дальнейшем предполагается приписывать связям дифференцированные числовые оценки, зависящие от контекста, причем это будет делаться непосредственно в правилах, устанавливающих связи (такой подход принят в системе О. С. Кулагиной).

Приведем некоторые данные о компьютерной реализации фрагментного синтаксического анализа. Собственно модуль анализа - это программа объемом около 2500 строк, написанная на языке PL/1, причем основная, управляющая часть алгоритма занимает в ней относительно небольшое место. Имеется также транслятор синтагм (около 1300 строк на PL/1), переводящий их из исходной формы в машинную.

По существу, синтагмы являются небольшими программами, записанными в виде графов переходов, и модуль анализа выполняет по отношению к ним функции интерпретатора. По своему лингвистическому содержанию они достаточно близки к синтагмам, применяемым в фильтровом алгоритме, поскольку описывают тот же класс СинтС. Однако между ними есть и различия, порождаемые такими особенностями модуля частичного фрагментного анализа, как: а) процедурная форма представления синтагм (в фильтровой системе используется декларативная форма); б) возможность оперировать признаками фрагментов; в) необходимость выделять ситуации, в которых устанавливаемые связи являются высоковероятными.

4. 8. 2. 2. Алгоритм установления связей

Ниже описывается алгоритм, устанавливающий высоковероятные синтаксические связи между словами фразы.

Алгоритм применяется после того, как произведен морфологический анализ, т. е. для каждого слова фразы получены все варианты его лексико-морфологического разбора – омонимы. Напомним, что они состоят из имени лексемы и набора морфологических показателей (например, слово "мыло" имеет три омонима: МЫТЬ, V, несов, изъяв, прош, ед, сред; МЫЛО, S, ед, им; МЫЛО, S, ед, вин).

Фрагментом будем называть группу омонимов, расположенных на нескольких последовательных местах фразы (по одному на каждом месте), с заданным на этих омонимах ориентированным

деревом синтаксических связей. Дуги дерева помечены именами СинтО. Отдельные омонимы также считаются фрагментами ("дерево" из одной вершины).

Объект, с которым работает алгоритм, - растущая последовательность фрагментов А, в каждый момент времени содержащая часть омонимов фразы и некоторые фрагменты, образованные из этих омонимов. Алгоритм, двигаясь от начала к концу последовательности, пытается связать каждый содержащийся в ней фрагмент F с другими фрагментами (стоящими в последовательности раньше F). Фрагменты, возникшие в результате установления связей, помещаются в конец последовательности А в том порядке, в каком они образованы.

Фрагмент F, рассматриваемый алгоритмом в данный момент, будем называть активным. Активными становятся поочередно все фрагменты последовательности (в том числе и одноэлементные), при этом не происходит каких-либо возвратов или скачков.

Хотя при движении алгоритма вдоль последовательности А ее "хвост" отодвигается ввиду появления новых фрагментов, в конце концов возникает такая ситуация, когда активным является последний фрагмент А, и сделать следующий шаг невозможно. Тогда к последовательности добавляется очередной омоним фразы; он становится активным, после чего работа алгоритма продолжается. Когда опять возникает ситуация, в которой сделать следующий шаг невозможно, к последовательности добавляется еще один омоним и т. д. Если же в подобной ситуации оказывается, что все омонимы фразы исчерпаны, работа алгоритма заканчивается.

Омонимы добавляются к последовательности А в порядке их расположения во фразе слева направо (это существенно), а стоящие на одном и том же месте - в произвольном порядке (здесь порядок не влияет на получаемые результаты). В начальный момент последовательность содержит единственный элемент - один из омонимов, стоящих на первом месте фразы, и он же является активным.

Для каждого активного фрагмента F алгоритм выбирает в последовательности А его левых соседей - фрагменты, расположенные во фразе рядом с F слева. При поиске соседей учитывается, что в последовательности А фрагменты располагаются в порядке возрастания их правых границ (это связано с

порядком, в котором к A добавляются омонимы). Между соседями устанавливается отношение предпочтения: из двух фрагментов более предпочтительным считается имеющий бóльшую длину; фрагменты равной длины считаются равноценными.

Работа алгоритма с активным фрагментом F сводится к тому, что левые соседи F перебираются в порядке убывания их предпочтительности, и для каждого рассматриваемого соседа делается попытка связать его с F. Если для некоторого E - соседа F - это удалось, то дальнейшие попытки ограничиваются соседями F, равноценными E; менее предпочтительные соседи не рассматриваются.

Попытка связывания F с его левым соседом E делается следующим образом. Рассматриваются связи, соединяющие некоторый омоним X фрагмента E с самым правым омонимом Y фрагмента F. Между связями устанавливается отношение предпочтения: более предпочтительными считаются связи, имеющие меньшую длину, а одинаковые по длине считаются равноценными. Иными словами, более предпочтительны связи X ⟶ Y и X ⟵ Y с омонимами X, расположенными правее; связи с одним и тем же X равноценны.

Попытка связывания E и F состоит в том, что потенциально возможные связи между X и Y перебираются в порядке убывания их предпочтительности и для каждой связи делается обращение к соответствующей синтагме - правилу, устанавливающему данную связь. В синтагме проверяются содержащиеся в ней условия, и если они выполнены, E и F соединяются данной связью и возникший фрагмент помещается в конец последовательности A. После этого перебор ограничивается связями, равноценными данной; менее предпочтительные связи не расматриваются.

В некоторых случаях (а именно, когда фрагменту F присвоен некоторый вспомогательный признак "ревиз", и его вершина Y_1 не является в нем самым правым омонимом) рассматриваются также связи X ⟶ Y_1 и X ⟵ Y_1. Процедура их установления полностью совпадает с описанной выше.

Алгоритм предусматривает специальный механизм для анализа союзных сочинительных конструкций типа E ⟶ *и* ⟶ F. В этом случае связываются сразу три фрагмента: E, F и сочинительный союз. Подробное описание того, как это делается, приведено в [Митюшин, 1988].

Когда последовательность A построена, в ней выделяется

совокупность С максимальных фрагментов (фрагмент из А включается в С, если отрезок, занимаемый им во фразе, не является строгой частью отрезка, занимаемого каким-либо другим фрагментом из А). Совокупность С считается окончательным результатом работы алгоритма.

В статье [Митюшин, 1988] содержится также более формальное изложение описанного здесь алгоритма и приводятся аргументы в пользу принятых решений. Реализация алгоритма на компьютере почти полностью совпадает с версией, описанной в статье. Единственное важное отличие - возможность при некоторых условиях отменять приоритет более коротких связей и более длинных фрагментов, чтобы сделать перебор соответствующих вариантов полным. Для этого в синтагмах используются два специальных действия (отдельно для фрагментов и связей). Например, приоритет более коротких связей отменяется во многих случаях, когда связь идет слева направо и зависимым членом является существительное в родительном падеже, поскольку в русском языке эта ситуация представляет собой регулярный источник неоднозначности.

4.8.2.3. Синтагмы

В отличие от алгоритма установления связей, синтагмы в компьютерной системе фрагментного анализа реализованы существенно иначе, чем описано в [Митюшин, 1988]. Важным содержательным моментом является отказ от требования "бесконтекстности" и снятие всех ограничений на характер проверяемых условий. Изменена и форма представления: синтагма имеет вид не формулы исчисления предикатов, а графа переходов, что расширяет ее алгоритмические возможности.

Синтагма представляет собой следующее. Имеется ориентированный граф, в котором фиксирована единственная "входная" вершина; выделено некоторое подмножество "возвратных" вершин (возможно, пустое); дуги, выходящие из одной и той же вершины, упорядочены. В графе могут быть кратные дуги. На каждой дуге указан предикат, проверяемый при проходе по этой дуге, или действие, выполняемое при проходе по дуге. При предикатах возможен знак отрицания. Считается, что и предикаты, и действия вырабатывают значение истинности (для действий это всегда "истина").

Работа синтагмы состоит в движении от входной вершины

вдоль "истинных" дуг, пока это возможно. Если достигнута висячая вершина, процесс заканчивается. Для невисячей вершины рассматриваются в заданном порядке выходящие из нее дуги и делается переход вдоль первой из дуг, для которой значение предиката/действия есть "истина". Если все выходящие дуги имеют значение "ложь" и данная вершина является невозвратной, процесс заканчивается. В противном случае делается возврат в ту вершину V, из которой был совершен переход в данную (т.е. последнее продвижение по дуге отменяется), и рассматриваются выходящие из V дуги, которые следуют за "отмененной". Если мы находимся во входной вершине, процесс заканчивается.

Совокупность предикатов и действий выбирается так, чтобы с их помощью было удобно проверять свойства рассматриваемых фрагментов и выполнять необходимые операции. Эти предикаты и действия во многом похожи на предикаты, описанные в гл. 2. Предикаты проверяют разнообразные условия, относящиеся к омонимам в рассматриваемых фрагментах, синтаксическим связям между ними, признакам фрагментов. Как и в системе предикатов, описанной в гл. 2, некоторые предикаты и действия вводят новые переменные: для заданного омонима Z_1 в том же фрагменте подбирается омоним Z_2, определенным образом связанный с Z_1. Если нужного Z_2 не оказывается, в предикатах вырабатывается значение "ложь".

Основным действием, присутствующим в каждой синтагме, является link(r) - образование нового фрагмента путем соединения заданного омонима X одного фрагмента с вершиной Y другого, соседнего с ним, синтаксической связью с отношением r.

В процессе экспериментов совокупность используемых предикатов/действий непрерывно изменяется и пополняется; в настоящее время их в системе около 40. Некоторые предикаты комментируются в связи с примером синтагмы (см. ниже).

Синтагма вводится в компьютер в виде текста, где первые пять строк (не считая комментариев) - "шапка", а затем описываются дуги графа переходов, по одной в каждой строке. Будем называть элементами строки цепочки символов без пробелов, отделенные друг от друга одним или несколькими пробелами.

Первая строка в записи синтагмы состоит из двух элемен-

тов - номера синтагмы и ее мнемонического имени, где номер есть число от 1 до 9999. Во второй и третьей строках указывается, к каким частям речи могут принадлежать омонимы, соединяемые данной связью. Вторая строка, относящаяся к слову-хозяину, имеет первый элемент X, остальные - обозначения частей речи. Третья строка относится к слову-слуге и устроена так же, но вместо X пишется Y. В четвертой строке, задающей направление устанавливаемой связи, два элемента: символ D и одна из цифр 1, 2, 3, что соответственно означает "вправо", "влево" и "в обе стороны". В пятой строке, задающей требования к проективности устанавливаемой связи, также два элемента: N и одна из цифр 0, 1, 2, 3. Смысл этих цифр следующий: 0 - устанавливаемая связь должна быть проективной; 1 - связь вправо может быть непроективной; 2 - связь влево может быть непроективной; 3 - связь в обе стороны может быть непроективной.

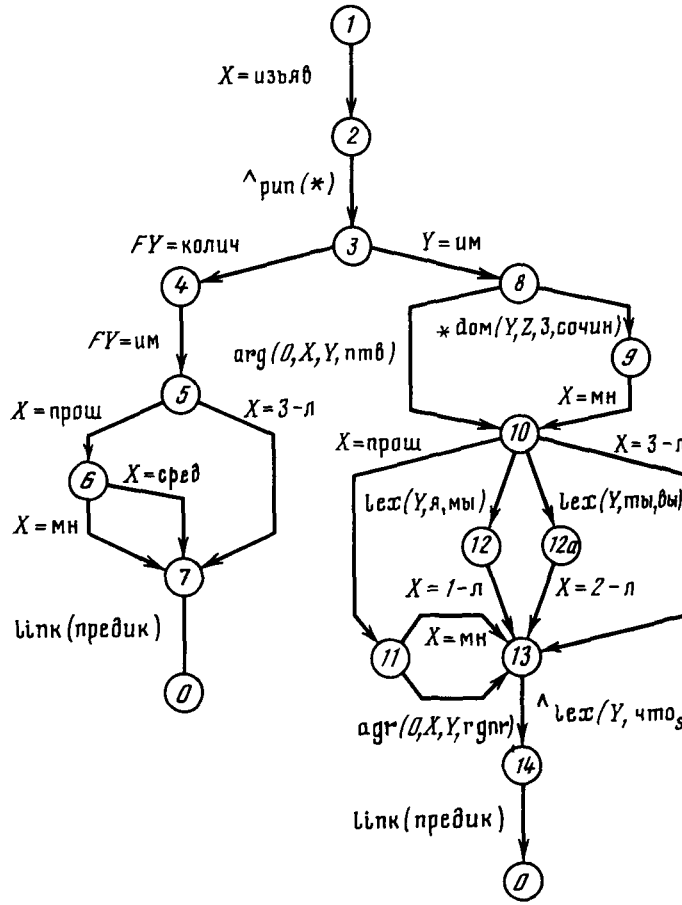
Далее идут строки, описывающие дуги графа. Предполагается, что висячие вершины помечены меткой 0, а невисячие - произвольными несовпадающими метками, отличными от 0. Метками могут быть любые цепочки символов, не содержащие пробела и знака *.

Первые два элемента строки, описывающей дугу, - метки начальной и конечной вершин, затем указывается имя предиката/действия и аргументы, разделенные пробелами. Скобки и запятые в качестве разделителей не используются. Перед предикатом может стоять знак отрицания ^.

Упорядочение дуг, выходящих из одной вершины, задается порядком их записи в синтагме. Входной вершиной считается начальная вершина первой дуги. В остальном порядок записи дуг произволен. Для возвратных вершин в одной из дуг, где такая вершина является начальной, справа к ее метке приписывается знак *. Отметим, что в реальных синтагмах возвратные вершины до сих пор появлялись очень редко.

В тексте синтагмы могут быть комментарии - произвольные строки, начинающиеся с символа /. При трансляции эти строки игнорируются.

Приведем пример синтагмы, заданной в виде графа переходов и в виде текста, предназначенного для ввода в компьютер. Эта синтагма имеет номер 1 и имя "предик.01". Она устанавливает предикативную связь от X к Y в случае, когда X



Синтагма "предик. 01"

- глагол в изъявительном наклонении, Y - существительное. Связь может быть направлена в обе стороны и всегда является проективной.

На рисунке синтагма изображена в виде графа переходов. Вершина 1 является входной, возвратных вершин нет. Дуги, выходящие из одной вершины, считаются упорядоченными слева направо (или против часовой стрелки).

Объясним смысл используемых предикатов. Символ = означает наличие у омонима или фрагмента данного признака. Так, X = изъяв означает, что X - глагол в изъявительном наклонении; FY = колич означает, что фрагмент, содержащий Y, есть количественная группа, и т.п. Предикат pun(t), где t -

мя знака препинания, истинен, если между рассматриваемыми фрагментами есть знак t; * означает "неопределенный" аргумент, т.е. pun(*) истинно, если между фрагментами есть какой-нибудь знак препинания.

Предикат agr(Ø,X,Y,c), где c - обозначение грамматической категории, означает согласование X и Y по этой категории. Он истинен, если у X и Y есть совпадающие характеристики данной категории, либо если у X или Y характеристики данной категории отсутствуют. Запись agr(1,X,Y,c) означает несколько более "жесткое" согласование: если у X есть характеристики данной категории, а у Y их нет, предикат имеет значение "ложь"; в остальном эти случаи совпадают. В примере встречаются категории числа (nmb) и рода (rgnr).

Предикат lex(Y, . .) означает, что Y является одной из перечисленных лексем. В текстовой форме синтагмы вместо имен лексем указываются их словарные номера.

Предикат *dom(Y,Z,3,сочин) находит для заданного омонима Y омоним Z, подчиненный Y по сочинительному отношению. Цифра указывает направление связи ("3" значит, что направление произвольно). В ситуациях неединственности берется самый левый из подходящих омонимов. Поскольку Z в дальнейшем не используется, предикат *dom(Y,Z,3,сочин) означает просто наличие у Y сочинительного "слуги".

Текстовая запись синтагмы имеет следующий вид:

```
1        предик. 01
/        самолет  <------  летит
X    V
Y    S
D    3
N    Ø
 1    2       =          X  изъяв
 2    3    ^  pun        *
 3    4       =          FY  колич
 3    8       =          Y  им
 4    5       =          FY  им
 5    6       =          X  прош
 5    7       =          X  3-л
 6    7       =          X  мн
 6    7       =          X  сред
 7    Ø       link       предик
```

```
 8   1Ø      agr Ø     X  Y  nmb
 8    9      *dom      Y  Z  3  сочин
 9   1Ø      =         X  мн
1Ø   11      =         X  прош
1Ø   12      lex       Y  488Ø  4877
/                        ( я )  ( мы )
1Ø   12a     lex       Y  4882  4883
/                        ( ты ) ( вы )
1Ø   13      =         X  3-л
11   13      agr Ø     X  Y  rgnr
11   13      =         X  мн
12   13      =         X  1-л
12a  13      =         X  2-л
13   14    ^ lex       Y  99
/                        ( что-S )
14    Ø      link      предик
```

Глава 5
ФОРМАЛЬНАЯ МОДЕЛЬ СЕМАНТИКИ

5.1. Возможные пути построения модели

В гл. 1 мы отмечали, что из трех компонентов ЛП - морфологического, синтаксического и семантического - достаточно высокой, если не исчерпывающей полноты можно добиться лишь в первых двух компонентах. Что касается семантики, то надо признать, что пока она не достигла такого уровня развития, при котором можно ожидать получения универсальных моделей. Мы имеем в виду семантические модели, охватывающие настолько широкий круг языковых явлений, что этого достаточно для решения любой прикладной задачи.

Впрочем, даже если бы такие модели уже существовали, это не решило бы одной существенной проблемы, связанной с настройкой модели на конкретную информационную систему (ИС). В самом деле, универсальная модель семантики имела бы на выходе семантический образ, записанный на универсальном формальном языке. Присоединение ЛП с такой семантикой к конкретной ИС потребовало бы создания алгоритмической процедуры, перерабатывающей универсальный семантический образ входного текста в некоторые данные этой ИС (и обратно). Конечно, эта процедура существенным образом должна быть настроена на способ представления данных в этой ИС. Таким образом, при данном взгляде на ЛП все сложности его настройки на конкретную предметную область и конкретный способ представления знаний выносятся в отдельные процедуры организации интерфейса ЛП ↔ ИС. Такие процедуры, по-видимому, оказались бы весьма непростыми.

По обеим указанным причинам мы с самого начала решили настроить семантическую модель нашего ЛП на конкретный способ представления данных в ИС[1]. Разумеется, теоретическая ценность ориентированной модели семантики ниже, чем универ-

---

[1] Подробнее об этом см [Boguslavskij, Tsinman, 1989; Богуславский, Цинман, 1990а, б, Boguslavskij, Tsinman, 1990].

сальной. Однако в практическом плане нас ожидает заметный выигрыш, поскольку такой шаг значительно облегчит процедуру семантического анализа и снимет проблему адаптации ЛП к конкретной ИС. Подчеркнем, что семантический компонент ЛП настраивается в первую очередь на способ представления данных в ИС и, в меньшей степени, на предметную область, из которой эти данные берутся. Это позволит переносить ЛП на другие предметные области (в рамках того же способа представления данных и той же общей задачи) без особых усилий.

При подобной постановке задачи важно выбрать такой способ представления данных, который:

1) достаточно широко распространен в стандартных пакетах программ для компьютеров разных типов;

2) достаточно универсален для того, чтобы с его помощью записывались данные из самых разных предметных областей;

3) обладает столь богатыми выразительными возможностями, что их полная реализация неосуществима посредством каких-либо облегченных диалоговых средств (графических или меню) и требует специального формального языка.

Существенно для нашего решения то обстоятельство, что способ представления данных для нас первичен по отношению к ЛП: берется готовый, широко распространенный способ представления данных, и ЛП настраивается на него. Это отличает наш подход от таких разработок, в которых первичным оказывается ЛП. В этом случае он применим только к таким ИС, которые построены специально для него.

В качестве ИС, которую должен обслуживать ЛП, мы выбрали реляционную базу данных с формальным языком запросов SQL. Перед ЛП ставится следующая задача: преобразовать запрос к базе данных, сформулированный на естественном языке, в запрос на языке SQL с тем, чтобы СУБД могла непосредственно реагировать на этот запрос. Это позволит продемонстрировать возможности ЛП в области понимания текстов - главной задачи, для которой он предназначен.

Приведем некоторые соображения, обосновывающие указанный выбор ИС и целесообразность поставленной перед ЛП задачи.

5.2. Типы СУБД и способы общения с ними

В настоящее время системы управления базами данных стали непременной составной частью математического обеспе-

жения любого компьютера. Возникла новая разветвленная ветвь компьютерной науки - теория баз данных. В ней рассматриваются три основные модели организации данных в СУБД: реляционная, сетевая и иерархическая. Не вдаваясь в сравнительный анализ этих моделей (см. об этом [Ульман, 1983]), отметим, что наиболее перспективной из них является реляционная. Коротко говоря, реляционная модель обладает дескриптивной мощностью двух других моделей при минимальном числе базисных понятий.

В реляционной СУБД все данные размещены в таблицах. Для каждой из них указан список атрибутов (имен столбцов), а для каждого атрибута задан перечень его возможных значений. Этими весьма экономными сведениями исчерпываются все знания о предметной области, с которыми должен оперировать ЛП.

Для общения с СУБД, конечно, никакие диалоговые средства типа меню недостаточны. Поэтому каждая СУБД снабжается специальным формальным языком запросов, с помощью которого пользователи могут получать справки о данных, хранящихся в СУБД. В теории баз данных определено понятие полноты языка запросов для реляционных СУБД. Язык запросов обладает полнотой, если в нем могут быть сформулированы запросы нескольких определенных категорий (формулировки теорем о полноте языка даются в терминах алгебраических операций над реляционными таблицами). Все реляционные СУБД должны снабжаться языком запросов, обладающим полнотой. Поэтому в некотором смысле все языки запросов, используемые в СУБД, эквивалентны. Тем не менее в последнее время наиболее популярным в реляционных СУБД становится язык запросов SQL (Structured Query Language), разработанный фирмой IBM в 1976 г. В настоящее время возникла целая индустрия ориентированных на SQL СУБД для больших, средних и малых компьютеров. Похоже, что SQL становится стандартным языком запросов для реляционных СУБД. Возможно, это связано с тем обстоятельством, что SQL обладает ясной логической природой и четкой структурой, в то время как другие языки запросов очень алгебраичны. Это становится определенным достоинством и с точки зрения ЛП, на выходе которого должно возникнуть выражение на языке запросов.

Язык SQL функционирует не только как язык запросов, но и как инструмент для ведения (пополнения и коррекции) данных

в базе. Более того, фирма IBM планирует превратить SQL в язык манипулирования данными в операционных системах, созданных этой фирмой; см. [Эршил, 1988].

Надо, однако, признать, что при всех указанных достоинствах SQL остается достаточно сложным формальным языком, требующим значительных усилий для овладения и пользования им. Отдавая себе в этом отчет, разработчики фирмы IBM в последнее время стали снабжать СУБД наряду с языком запросов SQL еще одним языком-посредником - QBE (Query-by-Example) [Sordi, 1984] Язык QBE выводит на экран дисплея шапки реляционных таблиц, содержащих данные, и предлагает пользователю фиксировать в соответствующих колонках требуемые значения атрибутов и указать сокращенные команды. Затем результат этой операции алгоритмически переводится в запрос на SQL. Простейшие запросы на QBE может задать и неподготовленный пользователь. Однако следует отметить, что, с одной стороны, язык QBE не обладает полнотой (в указанном смысле), а с другой стороны, в случае нетривиальных запросов требует от пользователя определенных навыков. Как признают сами разработчики [Kahn, 1984, p 107], язык-посредник QBE не смог существенно превзойти SQL в облегчении процедуры общения с СУБД для массового пользователя.

Таким образом, если стоит задача максимально облегчить доступ к информации, хранящейся в базах данных, широкому кругу пользователей, не знакомых ни со специальным языком запросов, ни со способом организации данных в базе, - а такая задача приобретает сейчас все большую актуальность, - то необходимо создать "технологию, позволяющую работать с менее структурированными языками запросов, которые называются в литературе естественными языками" [Kahn, 1984, p. 111]. Общая тенденция развития средств интерфейса - в сторону все большего их приближения к естественным языкам - обозначена здесь как нельзя более ясно. Для разработчиков ЛП важно быть на высоте требований этой новой тенденции.

Необходимо отметить, что работы, направленные на создание естественно-языкового интерфейса с базами данных, идут в двух направлениях. Представители первого направления стремятся обеспечить полноценное понимание запроса, не прибегая к серьезному лингвистическому анализу текста. Примером такого подхода в нашей стране может служить система

INTERBASE; см. [Диненберг и др., 1990]. Приверженцы второго подхода - к ним мы относим и себя - исходят из того, что даже в узкой предметной области вполне реальны запросы, имеющие настолько сложное содержание, что его адекватное понимание требует глубокого синтаксического и семантического анализа. Этот подход реализуется в ряде зарубежных разработок [Ginsparg, 1983; Lehmann et al , 1985; Maruyama, Watanabe, 1987; Maruyama et al , 1988] Мы надеемся, что материал, который будет представлен ниже (см., в частности, образцы запросов в разд. 7.1), подтвердит справедливость такой установки.

5.3. Язык семантических структур

5.3.1. Общие замечания

Как уже было сказано, целью семантического анализа является переработка синтаксической структуры запроса в семантическую структуру. Зачем это нужно? Ведь уже при синтаксическом анализе используется большой объем семантических сведений, благодаря чему СинтС может обеспечивать достаточно хороший контроль смысла. Именно это обстоятельство позволило получать в системах ЭТАП-1 и ЭТАП-2 переводы высокого качества. Почему же в ЛП нельзя ограничиться синтаксическим анализом?

Дело здесь в том, что при переводе с одного языка на другой (неважно, естественный или искусственный) глубина анализа определяется мерой структурного расхождения между языками. Языки европейского ареала и русский достаточно близки по своей структуре, чтобы при переводе можно было анализировать текст до уровня СинтС (точнее, до уровня нормализованной СинтС). В случае же ЛП расхождения между естественным языком и SQL настолько значительны, что для их преодоления оказалось необходимым спуститься на следующий уровень представления высказывания - семантический.

Как уже говорилось (см. разд. 4.2), СинтС являет собой дерево зависимостей, в узлах которого стоят лексемы, входящие в состав предложения, с их морфологическими характеристиками, а ветви помечены именами синтаксических отношений, специфичных для данного естественного языка. С формальной точки зрения СемС представляет собой объект того же вида - дерево зависимостей (подробнее см. ниже). Однако содержа-

тельный замысел у этих двух структур совершенно разный. Синтаксическая структура "обращена лицом" в сторону естественного языка. Она описывает синтаксическое строение конкретного русского предложения. Ее интересует, какие русские слова употреблены в этом предложении и как они связаны между собой. Поэтому в ней фигурируют имена лексем и специфические для русского языка синтаксические отношения. Семантическая структура, напротив, ориентирована в сторону предметной области. Ей безразлично, как было построено исходное предложение и даже на каком естественном языке оно было построено. Зато для нее важно, какую предметную область ей предстоит описывать, т. е. какие элементарные объекты в ней представлены и в какие связи они вступают.

Именно поэтому (о чем мы уже говорили в гл. 1) предыдущие компоненты ЛП (морфологический и синтаксический) могли разрабатываться безотносительно к предметной области, а при работе над семантическим компонентом ее пришлось сразу же фиксировать. Для экспериментальной версии ЛП предметной областью послужила демонстрационная база данных, функционирующая на основе СУБД ORACLE. Эта кадровая база данных содержит сведения о служащих некоторой условной фирмы, размещенные в четырех таблицах: EM (служащие), DP (отделы), CT (города) и ST (штаты). Перечислим некоторые атрибуты (столбцы) этих таблиц. В первой таблице есть атрибуты ENAME (фамилия служащего), JOB (должность), SAL (зарплата), DEPTNO (номер отдела), SUBORDIN (подчиненные данного служащего, если они есть) и др. Атрибуты второй таблицы: DEPTNO (номер отдела), DNAME (название отдела), LOC (город, в котором отдел расположен), MGR (фамилия начальника отдела). В третьей таблице имеется два атрибута: CITY (название города) и STATE (название штата, в котором расположен город). Четвертая таблица содержит всего один атрибут - STATE (название штата).

5.3.2. Запись запросов на языке SQL

Чтобы объяснить, какие требования разумно предъявить к СемС, следует дать более детальное представление о конечном результате, к которому стремится ЛП, - о записи запроса на языке SQL.

Стандартный запрос на языке SQL состоит из трех частей

строк). В строке SELECT указываются имена одного или нескольких атрибутов, значение которых составляет ответ на запрос. Можно запрашивать также некоторые функции от атрибутов - среднюю зарплату, количество служащих и т.п. В строке FROM сообщается, какую таблицу (или какие таблицы) следует рассматривать для получения ответа на запрос. Строка WHERE содержит условия, которым должны удовлетворять значения искомых или каких-либо других атрибутов таблиц, указанных в строке FROM. Эти условия записываются в виде выражений некоторого формального языка. Частями этих выражений могут быть другие SQL-предложения. Глубина "встроенности" SQL-предложений произвольна.

Укажем в качестве иллюстрации два запроса на естественном языке, относящиеся к выбранной предметной области, и их запись на языке SQL. Первый из этих примеров, более простой, мы уже приводили в гл. 1. Здесь мы его повторим для удобства сопоставления с более сложным вторым примером.

(1) *Каков номер отдела сбыта?*

```
(1') SELECT   DEPTNO
     FROM     DP
     WHERE    DNAME = 'сбыт'
```

(2) *Кто в коммерческом отделе получает самую высокую зарплату?*

```
(2')SELECT ENAME
    FROM   EM
    WHERE  DEPTNO = (SELECT   DEPTNO
                     FROM     DP
                     WHERE    DNAME = 'коммерческий')
       AND SAL = (SELECT MAX(SAL)
                  FROM   EM
                  WHERE  DEPTNO = SELECT   DEPTNO
                                  FROM     DP
                                  WHERE    DNAME = 'ком-
                                             мерческий')
```

5.3.3. Требования к семантической структуре (СемС)

Как видно по этим примерам, SQL-предложение имеет крайне мало общего с соответствующим ему выражением на естественном языке. Поэтому при построении ЛП принципиально важно найти удобный промежуточный уровень (который мы и назвали

уровнем семантических структур), обладающий следующими свойствами.

С одной стороны, СемС должны быть достижимы из входных ЕЯ-запросов (точнее, из их синтаксических структур) с помощью естественных лингвистических преобразований. С другой стороны, они должны допускать перевод на SQL посредством простой алгоритмической процедуры.

Таким образом, вид СемС определяется требованиями двоякого рода. Требования первого рода вытекают из особенностей кодирования информации в ЕЯ. Требования второго рода вытекают из особенностей записи информации на формальном языке запросов к БД.

В предложениях ЕЯ выражается весьма разнообразная информация, из которой для нашей задачи принципиальный интерес представляют два типа: 1) собственно семантическая информация, т.е. информация об объектах данной предметной области, их свойствах и связях между ними; 2) кореферентная информация, т.е. информация о том, в каких случаях речь идет об одном и том же объекте, а в каких - о разных.

В формальном семантическом образе запроса должна быть полностью отражена информация обоих типов. Однако поскольку желательно сохранить древесный характер СемС, в нее непосредственно включается только информация первого типа, а сведения о кореферентности задаются отдельно от нее.

С другой стороны, требования выразимости СемС в SQL также накладывают ряд ограничений на вид СемС. Ниже одно из таких ограничений будет проиллюстрировано.

Как известно, одна из основных трудностей семантического анализа связана с тем, что в естественном языке очень большую роль играют синонимия и омонимия: один и тот же смысл может быть выражен по-разному (синонимия), а одно и то же выражение может иметь много разных значений (омонимия). От языка семантических структур естественно ожидать, чтобы в нем синонимия и омонимия были сведены к минимуму. По существу, весь процесс семантического анализа можно представлять себе как борьбу с этими двумя явлениями: в каждый момент необходимо разрешать омонимию (то есть определять, в каком значении реализовано данное выражение) и снимать синонимию (то есть, заменять данное выражение на синонимичное ему, но более простое выражение).

С точки зрения требований, предъявляемых к семантическому языку, синонимия и омонимия не равноправны. Отсутствие омонимии является абсолютным требованием: если семантическую структуру можно понимать несколькими способами, то каждую интерпретацию необходимо фиксировать, и тогда именно ее и следует считать семантической структурой, которая тем самым станет однозначной.

С синонимией дело обстоит иначе. Наличие синонимии не столь противопоказано семантическому языку, как наличие омонимии. Необходимо лишь, чтобы синонимичные структуры были в каком-то смысле эквивалентны, например, вызывали тождественную реакцию воспринимающей их системы. Фактически каждый достаточно богатый формальный язык обладает определенной гибкостью и тем самым неизбежно в какой-то степени допускает синонимию. В некоторых пределах синонимия есть и в языке SQL. Один и тот же запрос в ряде случаев можно сформулировать двояко: посредством объединения таблиц или с помощью вложенных запросов. В такой же степени возможны и синонимичные семантические структуры. Забегая вперед, заметим, что в семантических структурах синонимия обусловлена эквивалентностью семантического отношения со значением 'такой, что' и конъюнктивного семантического отношения: 'зарплата служащего, такого, что его фамилия есть "Джоунз", равна 3000' = 'зарплата служащего равна 3000, и его фамилия есть "Джоунз"'.

5.3.4. Формальное определение СемС

Дадим теперь формальное определение объекта, который мы будем называть СемС запроса.

Как отмечалось выше, СемС является деревом зависимостей. Древесность - очень удобное свойство СемС, особенно потому, что входной объект семантического компонента ЛП - синтаксическая структура фразы - тоже является деревом зависимостей, и тем самым для работы с обоими видами структур может использоваться один и тот же аппарат алгоритмических преобразований.

5.3.4.1. Узлы СемС

В СемС узлы бывают четырех типов.

1. Термы - константы. Таковыми являются всевозможные значения атрибутов из таблиц ('Джоунз', 'менеджер',

'Чикаго',...), символы '?' и 'all', а также целые положительные числа из некоторого диапазона (для записи значений числовых атрибутов типа SAL).

2. Термы - переменные. К ним относятся имена таблиц; в нашей экспериментальной базе данных четыре таких терма: EM, DP, CT, и ST.

3. Функциональные символы. Функции - это атрибуты таблиц, такие как ENAME, SAL, DEPTNO, JOB, MGR, ... Для каждой функции указан список термов-констант, которые могут быть значениями этой функции, и указаны имена таблиц, для которых эта функция является атрибутом (т.е. заданы соответствующие данной функции термы-переменные).

4. Операторные символы. К ним относятся имена операторов языка SQL: MAX, MIN, AVG, SUM, COUNT, которые осуществляют действия над множеством значений атрибутов, затребованных в запросе. Для каждого оператора указывается список функций, над значениями которых этот оператор может производить действия.

Заметим, что во всех узлах СемС стоят элементарные фрагменты языка SQL, из которых должны конструироваться SQL-предложения.

5.3.4.2. Дуги СемС

Дуги СемС помечаются символами семантических отношений. Семантические отношения бывают четырех типов:

1. Аргументное отношение. Оно обозначается в СемС как arg.

2. Предикатные отношения: 'равно' (=), 'не равно' (≠), 'больше' (>), 'меньше' (<), 'больше или равно' (≥), 'меньше или равно' (≤), 'принадлежит' (∈), 'не принадлежит' (∉).

3. j-оператор. Он выражает отношение 'такой, что'.

4. Логические отношения конъюнкции (&) и дизъюнкции (v).

В примерах правил, которые будут приводиться ниже, семантические отношения будут иметь буквенные обозначения: EQU (=), NON-EQU (≠), MOR (>), LES (<), MOR-EQU (≥), LES-EQU (≤), JOT (j), CON (&) и DIS (v).

5.3.4.3. Поддеревья

СемС мы будем определять через семантические поддеревья. Мы различаем два вида поддеревьев: объектные и предикатные. Объектные поддеревья используются для записи на нашем се-

мантическом языке некоторых объектов или совокупностей объектов, а предикатные - для записи некоторых утверждений об объектах. Определение объектного поддерева ($\Delta_О$) и предикатного поддерева ($\Delta_П$) строится параллельно: сначала задаются элементарные $\Delta_О$ и $\Delta_П$, а затем - $\Delta_О$ и $\Delta_П$ произвольного вида.

В приводимых ниже определениях мы будем различать поддеревья, зависящие от переменной, и поддеревья, не зависящие от переменной. Поддеревья, не зависящие от переменной, мы будем называть константными.

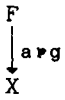
1. Элементарные объектные поддеревья

Выделяются три типа таких поддеревьев.

1. Вырожденное поддерево, состоящее из одного узла, в котором стоит терм-константа:

t

2. Если F - некоторый функциональный символ, а X - соответствующая ему переменная, то



- элементарное $\Delta_О$. Примеры:

Эти деревья читаются соответственно как 'фамилия служащего' и 'менеджер отдела'.

3. Если F - некоторый функциональный символ, X - соответствующая ему переменная, OPER - оператор, действия которого над множеством значений F допустимы, то

- элементарное Δ . Примеры:

(т.е. 'максимум зарплаты служащих', 'количество фамилий служащих').

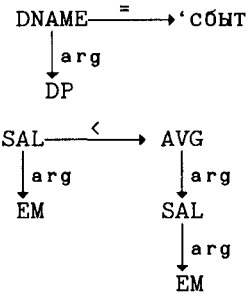
Элементарные Δ_0, определенные в п. 1, мы будем называть константными, а элементарные Δ_0, определенные в пп. 2 и 3, будем называть зависящими от переменной X и обозначать через $\Delta_0(X)$.

2. Элементарные предикатные поддеревья

Пусть Δ_0 и Δ_0^1 - элементарные объектные поддеревья. Тогда

$$\Delta_П = \Delta_0 \xrightarrow{r} \Delta_0^1,$$

где $r = \{=, \neq, <, >, \geqslant, \leqslant, \in, \notin\}$, - элементарное предикатное поддерево. При этом, если Δ_0 зависит от некоторой переменной X, то и результирующее поддерево $\Delta_П$ мы будем считать зависящим от той же переменной X. Примеры:



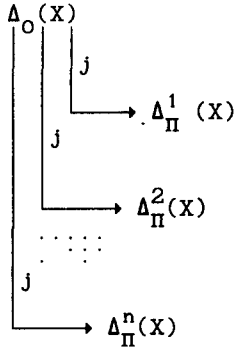
('название отдела - сбыт')

('зарплата служащего меньше средней зарплаты служащих')

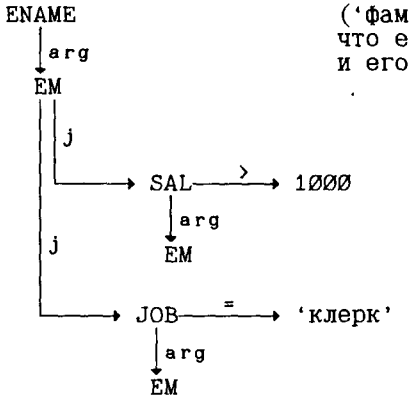
3. Объектные поддеревья

1. Всякое элементарное объектное поддерево является объектным поддеревом.

2. Пусть $\Delta_0(X)$ - объектное поддерево, а $\Delta_П^1(X)$, $\Delta_П^2(X)$, ..., $\Delta_П^n(X)$ - предикатные поддеревья, зависящие от одной и той же переменной X. Тогда



- объектное поддерево, зависящее от переменной X. Пример:



('фамилия служащего, такого, что его зарплата превышает 1000 и его должность - клерк')

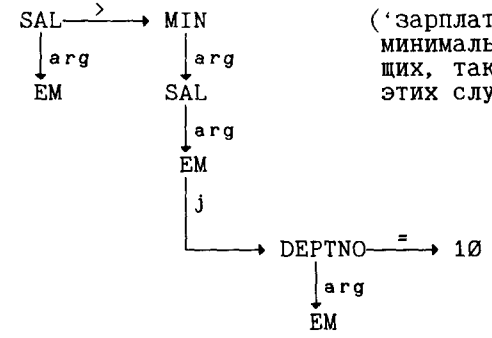
4. Предикатные поддеревья

1. Всякое элементарное предикатное поддерево является предикатным поддеревом.

2. Пусть Δ_O - элементарное объектное поддерево, а Δ_O^1 - какое-либо объектное поддерево. Тогда

$$\Delta_П = \Delta_O \xrightarrow{r} \Delta_O^1,$$

где $r \in \{=, \neq, <, >, \geqslant, \leqslant, \in, \notin\}$, - предикатное поддерево. При этом, если Δ_O зависит от некоторой переменной X, то и результирующее поддерево $\Delta_П$ мы будем считать зависящим от той же переменной X. Пример:

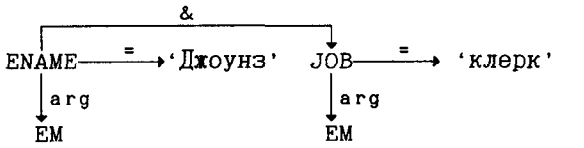


('зарплата служащего больше минимальной зарплаты служащих, таких что номер отдела этих служащих равен 10')

3. Пусть $\Delta_П^1$ и $\Delta_П^2$ - предикатные поддеревья. Тогда

$$\Delta_П = \Delta_П^1 \xrightarrow{r} \Delta_П^2,$$

где $r = \{\&, v\}$, - тоже предикатное поддерево. При этом, если $\Delta_П^1$ и $\Delta_П^2$ - не константные поддеревья, то они должны зависеть от одной переменной. В этом случае мы будем считать, что и результирующее поддерево $\Delta_П$ также зависит от этой переменной. Пример:



('фамилия служащего - Джоунз и должность служащего - клерк')

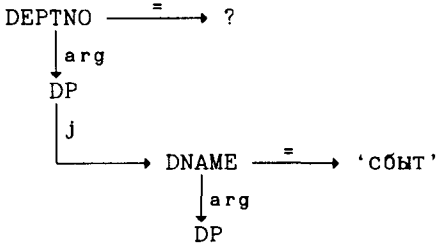
Теперь мы можем непосредственно определить СемС.

СемС - это предикатное (под)дерево, хотя бы один концевой узел которого есть '?'.

5.3.5. Примеры СемС

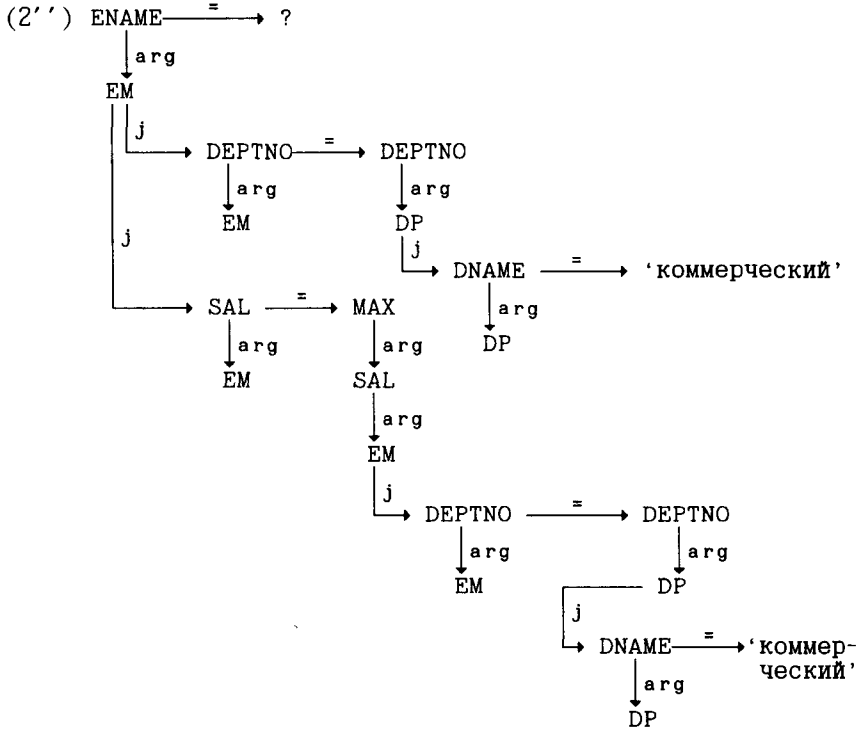
Приведем СемС рассмотренных выше запросов (1) и (2).

(1'')



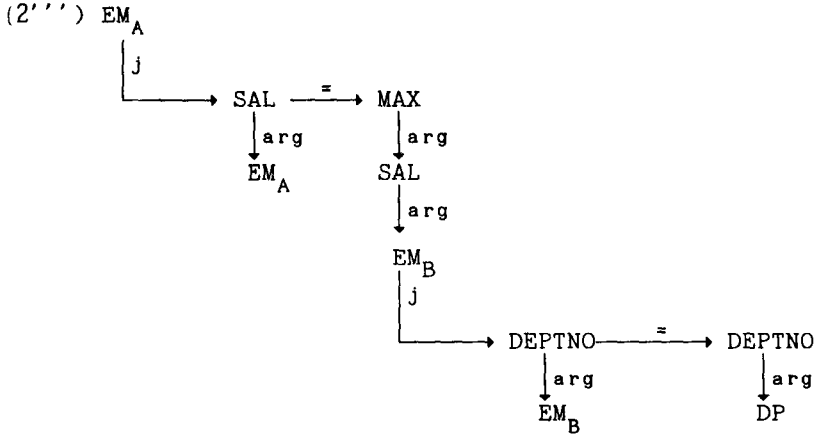
Прочтение СемС (1''):'Каков номер отдела, такого, что название этого отдела - 'сбыт'?'

(2'')



Прочтение СемС (2''): 'Каковы фамилии служащих, таких, что номер их отдела совпадает с номером отдела, название которого - 'коммерческий', а их зарплата равна максимальной зарплате служащих, таких, что номер их отдела совпадает с номером отдела, название которого - 'коммерческий'?'

Как указывалось выше, помимо самих СемС на выход семантического анализа подается информация о кореферентных связях в СемС. Эти связи, не отраженные непосредственно в СемС, могут устанавливаться между парами узлов СемС, в которых стоят одноименные термы-переменные. Кореферентная связь указывает на то, что оба вхождения переменной содержательно обозначают один и тот же объект. Так, в структуре (2'') некоторые вхождения переменной EM кореферентны, а некоторые нет. Приведем фрагмент (2''') структуры (2''), в котором эти сведения отражены с помощью дополнительных индексов:

(2''') 

В языке SQL имеется специальное средство, называемое синхронизацией, которое реализует эту информацию. Некоторые типы запросов невозможно записать на SQL без использования синхронизации.

Таким образом, мы построили достаточно строгое определение СемС как некоторого дерева зависимостей. В узлах этого дерева стоят элементы языка SQL, а его дуги помечены именами, однозначно интерпретируемыми в SQL. Определение СемС строилось с таким расчетом, чтобы СемС запросов, с одной стороны, и запись этих запросов на языке SQL, с другой стороны, были (в некотором смысле) изоморфны. Это заставило нас ввести некоторые ограничения на язык СемС, поскольку

SQL, при всем его богатстве, не позволяет выразить некоторые типы запросов. Чаще всего такие запросы можно обработать, только разбив их на два самостоятельных подзапроса. Например, язык SQL не позволяет построить единый запрос, соответствующий предложению ***В каком отделе средняя зарплата аналитиков превосходит максимальную зарплату клерков?*** Мы предусматриваем, что ЛП будет распознавать подобные случаи и сообщать пользователю о необходимости переформулировки или декомпозиции запроса.

Таким образом, каждая правильно построенная СемС может быть преобразована в запрос на языке SQL.

Можно с уверенностью утверждать, что информационные потребности массового пользователя могут быть вполне удовлетворены запросами, которые на семантическом уровне представимы в виде определенных выше СемС. Однако целый ряд запросов, ориентированных на некоторые вспомогательные возможности языка SQL, мы не готовы обрабатывать (т. е. не беремся получать для них СемС). Это в основном те возможности SQL, которые связаны не с поиском требуемой информации, а с дополнительным сервисом, обеспечивающим получение результата в форме, удобной для обзора: всевозможные сортировки выходных данных, организация их в виде наглядных таблиц, подготовка разнообразной отчетности. Дело в том, что такого рода справки могут интересовать лишь постоянных абонентов СУБД, в то время как наш ЛП рассчитан, в первую очередь, повторяем, на массовых, случайных пользователей.

Именно эта категория пользователей в максимальной степени нуждается в естественно-языковом интерфейсе, поскольку именно для них необходимость овладения специальным языком запросов становится реальным препятствием для обращения к СУБД. Такими пользователями являются, например, посетители выставки. У любознательных посетителей могут, как известно, возникать достаточно сложные вопросы, касающиеся сопоставительных характеристик изделий, которые представлены на стендах разных павильонов, фирм, городов, стран. Информационное обслуживание таких пользователей удобнее всего осуществлять на ЕЯ, т. е. с помощью ЛП описываемого типа. Постоянные же абоненты СУБД могут себе позволить овладеть языком запросов непосредственно или использовать квалифицированного оператора.

В первой версии ЛП мы не будем обрабатывать также запросы, требующие производства арифметических действий над значениями поисковых атрибутов. Потребность в такого рода подсчетах редко возникает в выбранной нами предметной области.

Внимательное рассмотрение приведенных выше примеров показывает, что переход от языка СемС к языку SQL представляет лишь некоторые технические трудности. По этой причине мы в дальнейшем не будем останавливаться на описании данного перехода, а сосредоточимся на гораздо менее тривиальном этапе получения СемС из СинтС запроса. Здесь возникает целый ряд проблем лингвистического характера. Этим проблемам посвящен следующий раздел.

5.4. Правила семантического анализа

Преобразование синтаксической структуры запроса в его семантическую структуру разбивается на три этапа:

1) нормализация СинтС;
2) семантизация нормализованной СинтС;
3) канонизация семантизованной СинтС.

Ниже мы ограничимся содержательным рассмотрением этих задач, оставляя формальные правила за кадром. Приводя примеры преобразований, мы будем стремиться избегать изображения формальных структур там, где это возможно и помогает повысить наглядность.

5.4.1. Нормализация СинтС

Синтаксическая структура запроса, которая была получена в ходе синтаксического анализа, неудобна для семантизации. Она сохраняет еще слишком много особенностей естественного языка, не имеющих прямого соответствия в языке семантических структур. Задача этапа нормализации состоит в том, чтобы устранить эти особенности. По существу, нормализованная структура - это то приближение к идеальной глубинно-синтаксической структуре, которое необходимо в рамках данной задачи (ср. в [Апресян и др., 1989, с. 125-150] близкое, но не тождественное содержание этапа нормализации в рамках задачи машинного перевода).

Конкретнее, этап нормализации должен превратить предложение в несколько более элементарных пропозициональных структур, допускающих прямой перевод в СемС. С этой целью СинтС подвергается таким преобразованиям, как опущение из-

быточных элементов, восстановление недостающих элементов и некоторые другие типы приведения структуры к виду, при котором она передает значение более прямо.

К числу семантически ненаполненных элементов, подлежащих элиминации на этом этапе, относятся, в частности, пустые предлоги в комплетивных, обстоятельственных, атрибутивных и элективных конструкциях (например, *работает (в) Чикаго <(в) отделе 20>, кто (из) клерков* и т.п.); полувспомогательные лексико-функциональные глаголы (*получать зарплату, занимать должность, носить имя*); пустые существительные в позиции хозяина (*размер зарплаты ⇒ зарплата*) или слуги (*зарплата в размере 3000 ⇒ зарплата 3000*); существительные с родовым значением в аппозитивной конструкции (*штат Алабама ⇒ Алабама*); связочные глаголы в присвязочных и комплетивных конструкциях (*является клерком, был равен*).

Обработка подобных неполнозначных элементов не всегда сводится к их опущению. Часто опущение должно сопровождаться определенными синтаксическими преобразованиями. Например, лексико-функциональные глаголы типа *получать (зарплату)*, как известно, являются хозяевами некоторых актантных связей смыслового существительного. Так, в предложении *Джонз получает зарплату 3000 долларов* никакие синтаксические связи слова *зарплата*, соответствующего функции SAL, не дают оснований для того, чтобы заполнить аргументную позицию этой функции. Это можно сделать, только введя в рассмотрение более широкий контекст и заимствовав подлежащее у глагола *получать*.

Обратим внимание на следующее обстоятельство: в правиле опущения пустых предлогов (*работает (в) Чикаго*) мы в первый, но не в последний раз сталкиваемся с тем, что правила семантического анализа зависят от описываемого мира. С точки зрения русского языка в обстоятельственном сочетании *работает в Чикаго* предлог *в* не является семантически пустым. В этой же конструкции могут оказаться и другие, несинонимичные ему предлоги, например, *работать под <около> Чикаго*. Поэтому опущение предлога в этой позиции, строго говоря, приводит к смысловой потере. Однако в том узком мире, который описывает наша база данных, все эти варианты отсутствуют. Работать можно только *в* Чикаго, и никак не *под* или *около*. Подобные соображения, возникающие при описании

любого ограниченного мира, естественно, упрощают семантический анализ и позволяют быстрее прийти к результирующей СемС.

Необходимо отметить, что не все виды семантической неполнозначности обрабатываются на этапе нормализации. Например, в разд. 5.4.2.1 мы будем обсуждать семантизацию глагола ИМЕТЬСЯ, который в большинстве случаев имеет непустой семантический образ, но в определенных условиях требует лишь синтаксической перестройки и не порождает никаких семантических элементов.

Среди элементов, присутствующих в СинтС лишь имплицитно, требуют восстановления следующие: субъекты в причастных и деепричастных оборотах *(отделы, расположенные в Чикаго ⇒ отделы, такие, что (эти)*[2] *отделы расположены в Чикаго)*; антецеденты анафорических и указательных местоимений *отделы, которые расположены в Чикаго ⇒ отделы, такие, что (эти) отделы расположены в Чикаго; фамилии клерков и их зарплата ⇒ фамилии клерков и зарплата (этих) клерков)*; опущенные определения при субстантивированных адъективах *(работающие в Чикаго ⇒ служащие, работающие в Чикаго)*; сокращенные элементы в сравнительных оборотах (*зарплата, как у Джоунза ⇒ зарплата, как зарплата у Джоунза; зарабатывает больше, чем Джоунз ⇒ зарабатывает больше, чем зарабатывает Джоунз; работает там же, где Смит ⇒ работает там же, где работает Смит*).

В одном ряду со сравнительными оборотами находятся сочинительные конструкции: те и другие содержат сокращенные элементы. Однако, в отличие от сравнительных конструкций, сочинительные в большинстве случаев не требуют развертывания - благодаря тому, что язык семантических структур располагает достаточными средствами изображения конъюнктивных и дизъюнктивных отношений. Впрочем, в некоторых ситуациях сочинительное развертывание целесообразно. Любопытно, что оно необходимо именно тогда, когда сочинительная конструкция находится в контексте сравнительной. Существенно, что

---

[2]Как отмечалось в разделе 5 3 3, в семантической структуре используются специальные средства для фиксации информации о кореферентности объектов Здесь и ниже в примерах мы будем для наглядности отражать эту информацию с помощью слова *этот*, заключенного в скобки

порядок развертывания этих двух сокращенных конструкций не произволен. Сочинительное развертывание должно предшествовать сравнительному:

(3) *должность и зарплата, как у Джоунза ⇒ должность, как у Джоунза, и зарплата, как у Джоунза ⇒ должность, как должность Джоунза, и зарплата, как зарплата Джоунза.*

Если бы разворачивание производилось в обратном порядке, т. е. сначала производилась обработка сравнительного оборота, а потом - сочинительного, мы бы имели

(3') *должность и зарплата, как у Джоунза ⇒ должность и зарплата, как должность и зарплата Джоунза,*

откуда было бы труднее получить дистрибутивную интерпретацию *должность, как должность Джоунза, и зарплата, как зарплата Джоунза.*

С проблемой сочинительного развертывания связаны еще два важных вопроса.

Во-первых, как известно, в сочинительных конструкциях часто имеется неоднозначность, обусловленная тем, относится ли некоторый элемент предложения ко всем сочиненным членам или только к одному из них. Например, запрос

(4) *Укажите фамилии аналитиков и коммивояжеров отдела 20*

можно понять двояко в зависимости от того, идет ли речь обо всех аналитиках или только об аналитиках отдела 20. Разрешить эту неоднозначность, как и многие другие типы синтаксической неоднозначности, можно только в интерактивном режиме работы ЛП. Пользователь должен иметь возможность контролировать правильность интерпретации своего запроса. Однако делать это, глядя на конечный результат работы ЛП - запрос на языке SQL, - крайне затруднительно, поскольку пользователь может не понимать этого языка. Поэтому в ЛП планируется специальный контрольный механизм. Анализируя запрос, ЛП выбирает одну из возможных интерпретаций, наиболее предпочтительную, но построив для этой интерпретации СемС, ЛП будет подавать ее на вход синтезирующей процедуры. Эта процедура должна восстановить по СемС запрос на русском языке, не допускающий двоякого понимания. Если пользователь признает этот запрос синонимичным своему исходному замыслу, СемС подается на дальнейшую переработку. Если же он имел в виду другую интерпретацию, то он может переформулировать свой запрос так, чтобы сделать его более однозначным. Так,

например, при обработке сочинительных конструкций типа (4) выбирается интерпретация, при которой спорный элемент относится ко всем сочиненным членам. Чтобы отразить другое понимание, пользователь получит возможность сформулировать запрос иначе, например, так:

(5) *Укажите фамилии коммивояжеров отдела 20 и всех аналитиков фирмы.*

Такой подход позволит избежать неправильной интерпретации во всех ситуациях, когда возможно неединственное понимание запроса (см., например, ниже правила семантизации операторных слов типа *максимальный*).

Второй вопрос, требующий обсуждения в связи с сочинительным развертыванием, касается проблемы дистрибутивности сочиненных групп. Запрос

(6) *Сколько клерков и менеджеров работает в Чикаго?*

можно понимать объединительно или дистрибутивно. При первом понимании речь идет о суммарном количестве служащих, занимающих в Чикаго должность клерка или менеджера. При втором понимании спрашивающего интересуют две цифры - отдельно по клеркам и по менеджерам. Наши правила семантического анализа ориентированы на понимание первого типа. Если пользователь обнаружит в интерактивном режиме, что ему нужна вторая интерпретация, то ему придется переформулировать свой запрос, например, так: *Сколько клерков и сколько менеджеров работает в Чикаго?*

Помимо устранения семантически ненаполненных и восстановления имплицитных элементов, на этапе нормализации проводятся и некоторые другие канонизирующие преобразования, из которых мы отметим два.

Во-первых, вопросительное слово или сочетание переносится в вершину предложения с образованием придаточного относительного. Например:

(7а) *Зарплата каких клерков заключена в интервале от 2000 до 3000? ⇒ Каковы клерки (такие, что) зарплата (этих) клерков заключена в интервале от 2000 до 3000.*

(7б) *В каких отделах средняя зарплата служащих не превышает 3000? ⇒ Каковы отделы (такие, что) средняя зарплата служащих в (этих) отделах не превышает 3000?*

Такая перестройка структуры продиктована стремлением максимально упростить переход от СемС к SQL-выражению. Как

было видно из примеров СемС и соответствующих им SQL-запросов, которые приводились в разд. 5.3.2, вершинное положение вопросительного элемента в СемС приводит к тому, что СемС "разворачивается" совершенно параллельно тому, как "разворачивается" SQL-выражение.

Во-вторых, производится восстановление грамматически исходной формы глагола: замена пассива на актив. Например:

(8) *Какой отдел возглавляется Джоунзом? ⇒ Какой отдел возглавляет Джоунз?*

После этих преобразований запрос подается на этап семантизации.

5.4.2. Семантизация нормализованной СинтС

На вход этапа семантизации поступает структура, в которой устранены многие языковые средства, неявно выражающие значение, отсутствуют многие семантически незначащие слова, но еще не появилось ни одного элемента семантического языка. Задача этого этапа - центрального во всем процессе перехода СинтС ⇒ СемС - состоит в том, чтобы истолковать все элементы полученной структуры в терминах языка семантических структур. В нормализованной СинтС в таком истолковании нуждаются элементы двух типов - лексемы и синтаксические отношения. Семантически содержательные элементы третьего типа, которые присутствуют в нормализованной СинтС, - морфологические характеристики при лексемах - в силу особенностей предметной области не должны подвергаться семантизации. Каковы эти особенности?

Семантически содержательные морфологические характеристики имеются у глаголов и у имен. Глагольные морфологические характеристики (время, вид и наклонение) не актуальны, потому что описываемый нами мир статичен и не развивается во времени. Конечно, в этом мире могут происходить изменения, которые приводят к изменению состояния базы данных (например, служащие поступают на работу и увольняются, переходят из отдела в отдел, получают повышения и т.п.), но каждый раз, когда мы обращаемся к базе данных с запросом, мы имеем дело с некоторым фиксированным состоянием дел. В этом мире существуют только ситуации типа "зарабатывает", но не "зарабатывал раньше" или "будет зарабатывать" или "зарабатывал бы".

Единственная семантически содержательная характеристика имен - характеристика числа - не должна семантизироваться со другой причине. Во-первых, в части контекстов и эта характеристика не является семантической, например, тогда, когда при существительном имеется числительное (*пять отделов*) или существительное стоит в предикативной позиции (*работает <является> клерком*). Во-вторых, даже в тех случаях, когда эта характеристика действительно является смыслоразличительной (ср. *какой город ≠ какие города*), язык SQL не реагирует на это различие. Это значит, что на запрос типа *В каком городе средняя зарплата служащих превышает 3000?* в качестве ответа предъявляются все города, обладающие указанным свойством, т. е. запрос интерпретируется так, как если бы существительное *город* стояло во множественном числе.

Все это позволяет нам пока полностью отвлечься от проблемы семантизации морфологических характеристик. Впрочем, при развитии системы в сторону способности к кооперативному диалогу проблема морфологических характеристик встанет вновь. Дело в том, что морфологические характеристики могут служить средством выражения пресуппозиций, а без их учета кооперативный диалог невозможен. Например, запрос

(9) *В каком отделе работают Кларк и Джоунз?*

предполагает, что названные лица работают в одном и том же отделе, в то время как запрос

(10) *В каких отделах работают Кларк и Джоунз?*

такого предположения не содержит. Система, ориентированная на кооперативный диалог, должна уметь обнаруживать ситуации нарушения пресуппозиций и соответствующим образом на них реагировать.

Этап семантизации состоит из трех групп правил: 1) семантизация лексики; 2) семантизация синтаксических отношений; 3) оптимизация, или канонизация, семантической структуры. Рассмотрим эти группы в указанном порядке.

5.4.2.1. Семантизация лексики

Слова, фигурирующие в нормализованной СинтС и требующие истолкования, удобно классифицировать по тому, как они соотносятся с единицами семантического языка. С этой точки зрения они делятся на две группы. В первую группу попадают

слова, имеющие в семантическом языке прямые аналоги. Такие аналоги есть у всех четырех типов узлов СемС - функций, переменных, констант и операторов, - а также у предикатных семантических отношений. Вторую группу образуют слова, не имеющие в семантическом языке прямых аналогов. Начнем со слов первой группы.

Функции семантического языка выражаются существительными, глаголами и глагольно-субстантивными сочетаниями. Семантизация этих слов производится с помощью правил-толкований, в задачу которых, как это будет видно из примеров, помимо собственно истолкования слова, входит также правильное размещение в СемС слов, синтаксически связанных с толкуемым словом. (Здесь и ниже в примерах семантический образ приписывается только толкуемому элементу; остальные элементы примера остаются в неизменном виде и должны истолковываться другими правилами.)

(11) *Джоунз зарабатывает 5000* ⇒ SAL —(=)→ 5000
arg ↓ *Джоунз*

Джоунз руководит бухгалтерией ⇒ MGR —(=)→ *Джоунз*
arg ↓ *бухгалтерия*

должность Джоунза ⇒ JOB
arg ↓ *Джоунз*

должность менеджера ⇒ JOB —(=)→ *менеджер*

работать аналитиком ⇒ JOB —(=)→ *аналитик*

местожительство Смита ⇒ LOC
arg ↓ *Смит*

фамилия служащего ⇒ ENAME
arg ↓ *служащий*

Во многих случаях слово может соответствовать разным функциям в зависимости от контекста. Например:

(12) *Джоунз работает в отделе 10* ⇒ DEPTNO —(=)→ 10
arg ↓ *Джоунз*

Джоунз работает в Чикаго ⇒ LOC —(=)→ *Чикаго*; LOC —(arg)→ *Джоунз*

Джоунз работает в Калифорнии ⇒ ST —(=)→ *Калифорния*; ST —(arg)→ *Джоунз*

13) *название отдела* ⇒ DNAME —(arg)→ *отдел*

название города ⇒ CITY —(arg)→ *город*

Следующая - самая немногочисленная - группа состоит из слов, соответствующих переменным семантического языка. Это слова СЛУЖАЩИЙ <ЛИЦО, СОТРУДНИК> (EM), ОТДЕЛ <ПОДРАЗДЕЛЕНИЕ> (DP), ГОРОД (CT) и ШТАТ (ST).

Еще один тип образуют слова, обозначающие константы. Это такие слова, как ДЖОУНЗ, СМИТ, БУШ и т.п. (значения функции ENAME), МЕНЕДЖЕР, КЛЕРК, АНАЛИТИК и т.п. (значения функции JOB), ЧИКАГО, ДАЛЛАС, ВАШИНГТОН и т.п. (значения функции LOC), КАЛИФОРНИЯ, НЬЮ-ДЖЕРСИ, ТЕХАС и т.п. (значения функции ST), БУХГАЛТЕРИЯ, СБЫТ, ИССЛЕДОВАТЕЛЬСКИЙ и т.п. (значения функции DNAME). Следует учитывать, что такие слова имеют два типа употреблений - автонимное (константное) и неавтонимное (неконстантное).

Употребление первого типа предполагает, что при семантизации слово просто заменяется на соответствующую константу:

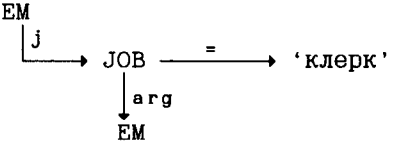
(14) *Джоунз* ⇒ 'Джоунз'
клерк ⇒ 'клерк'
бухгалтерия ⇒ 'бухгалтерия'
Чикаго ⇒ 'Чикаго'
Калифорния ⇒ 'Калифорния'.

Употребление второго типа требует истолкования слова в терминах функций и констант:

(15) *Джоунз* ⇒ EM —(j)→ ENAME —(=)→ 'Джоунз'; ENAME —(arg)→ EM

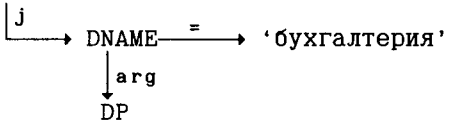
'служащий, фамилия которого - Джоунз';

клерк ⇒ EM



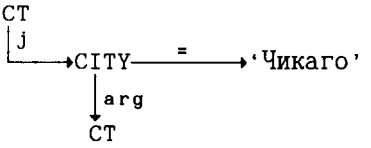
'служащий, должность которого - клерк';

бухгалтерия ⇒ DP



'отдел, название которого - бухгалтерия';

Чикаго ⇒ CT



'город, название которого - Чикаго'.

Мы уже говорили о том, что одна из главных трудностей при анализе текста состоит в широком распространении явления омонимии разного рода. В разд. 3.1 приводились примеры морфологической и лексико-грамматической омонимии, т.е. ситуации, когда одна и та же словоформа имеет альтернативные наборы морфологических характеристик и/или может быть отнесена к разным лексемам. Различие автонимного и неавтонимного типа употребления слова, с которым мы имеем дело здесь, представляет собой еще одно явление того же ряда: для правильного семантического анализа данного слова необходимо определить тип его употребления. Это можно сделать по контексту. Автонимное употребление имеет место в следующих двух классах контекстов.

А. Если данное слово связано отношением "=" с соответствующей ему функцией, а именно:

а) существительное с признаками "человек" и "собст" связано с функцией ENAME, восходящей к словам ИМЯ, ФАМИЛИЯ, ЗВАТЬ и т.п., или с функцией MGR, восходящей к словам ВОЗГЛАВЛЯТЬ, РУКОВОДИТЬ и т.п.; например: ***Какова должность служащего по фамилии Джоунз?***; ***Где работает клерк, которого зовут Джоунз?***; ***Какой отдел возглавляет Кларк?***;

б) существительное с признаком "должность" связано с функцией JOB, восходящей к словам ДОЛЖНОСТЬ, РАБОТАТЬ

кем), БЫТЬ (кем), ЯВЛЯТЬСЯ (кем) и т.п.; например: *Кто занимает должность клерка в отделе 10?*; *Назовите жителей Чикаго, работающих клерками*; *Назовите служащих, работающих в Чикаго в должности клерка*;

в) существительное или прилагательное с признаками "отдел" и "собст" связано с функцией DNAME, восходящей к словам НАЗЫВАТЬСЯ, НАЗВАНИЕ и т.п. Например: *Кто возглавляет отдел, расположенный в Чикаго и имеющий название "Сбыт"?*;

г) существительное с признаками "город" и "собст" связано с функцией LOC, восходящей к словам РАБОТАТЬ (в), ЖИТЬ в), ПРОЖИВАТЬ (в), ЖИТЕЛЬ и т.п. или с функцией CITY, восходящей к словам типа НАЗВАНИЕ, НАЗЫВАТЬСЯ и т.п.; например: *Какие клерки работают <живут> в Чикаго ?*; *Какие жители Чикаго зарабатывают больше 3000?*; *В каком штате расположен город под названием Чикаго?*

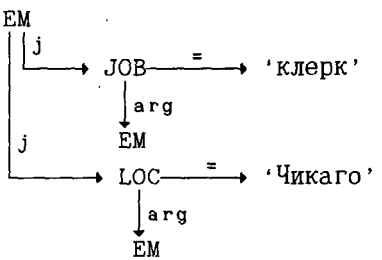
Б. Если данное слово связано отношением "&" и "v" с соответствующей константой (т.е. со словом, обладающим теми же признаками, но уже превратившимся в константу). Ср: *Кто занимает в отделе 20 должности клерков* (контекст А), *аналитиков* (контекст Б) и *менеджера?* (контекст Б).

В остальных контекстах слова указанных типов имеют неавтонимное употребление и должны толковаться. Например, сочетание

(16) *клерки, которые работают в Чикаго*

содержит автонимный контекст для слова *Чикаго* и неавтонимный для слова *клерки*:

(16′)



Специфическую группу составляют слова с вопросительным значением, т.е. такие, которые при семантизации порождают вопросительную константу '?'. К этой группе относятся как собственно вопросительные слова КТО, ЧТО, ГДЕ, СКОЛЬКО, КАКОЙ и т.п., так и глаголы типа УКАЗЫВАТЬ, ПЕРЕЧИСЛЯТЬ, НАЗЫВАТЬ, СООБЩАТЬ, которые в запросах употребляются в этой функции в форме инфинитива или повелительного наклонения.

При семантизации этих слов существенное значение имеет контекст, поскольку именно он определяет, с какой функцией связывается константа '?'. Например, слово КАКОЙ семантизируется следующим образом:

а) если оно относится к слову с функциональным значением (типа ЗАРПЛАТА), то его толкование исчерпывается константой '?':

(17) *какая зарплата* ⇒ SAL $\xrightarrow{=}$?
какова численность ⇒ COUNT $\xrightarrow{=}$?
каково место работы ⇒ DEPTNO $\xrightarrow{=}$?

б) если оно относится к слову константного типа, то, помимо символа '?', должно вырабатываться имя функции:

(18) *какие клерки* ⇒ ENAME $\xrightarrow{=}$?
↓arg
клерки

какой город ⇒ CITY $\xrightarrow{=}$?
↓arg
город

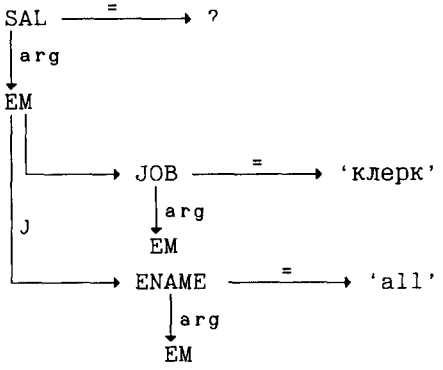
какой отдел ⇒ DEPTNO $\xrightarrow{=}$?
↓arg
отдел

Среди слов, соответствующих константам семантического языка, особняком стоит слово КАЖДЫЙ. Его семантическим образом служит константа 'all', которая, как и вопросительная константа '?', может присоединяться с помощью отношения '=' к любой функции. Это резко отличает 'all' и '?' от всех остальных констант, которые имеют очень избирательную сочетаемость с функциями (подробнее см. в разд. 6.2). Примеры:

(19) *Какова средняя зарплата в каждом отделе?*

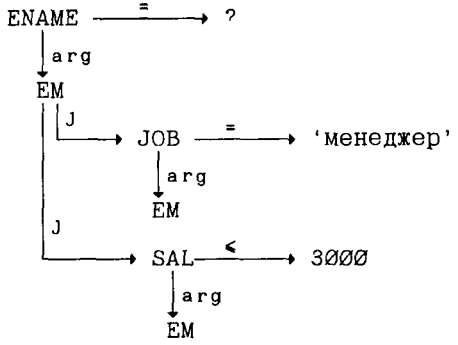
(19') AVG $\xrightarrow{=}$?
↓arg
SAL
↓arg
DP
↓J
DEPTNO $\xrightarrow{=}$ 'all'
↓arg
DP

20) *Для каждого клерка укажите его зарплату.*

20') 

Следующая группа слов соответствует предикатным отношениям семантического языка ('=', '≠', '>', '<', '≥', '≤', '∈', '∉'). Это такие слова, как РАВНЯТЬСЯ, РАВНЫЙ, ОТЛИЧНЫЙ (*от*), СОСТАВЛЯТЬ, ПРЕВОСХОДИТЬ, ПРЕВЫШАТЬ, УСТУПАТЬ; БОЛЬШЕ, МЕНЬШЕ, СВЫШЕ и т. п., а также их сочетания с отрицанием (*не равный, не превосходить, не превышать, не уступать* и т. п.). Например:

(21) *Зарплата каких менеджеров не превосходит 3000?*

(21') 

К этой же группе слов относится и сравнительный союз КАК, соответствующий предикату '=' и выступающий в запросах типа

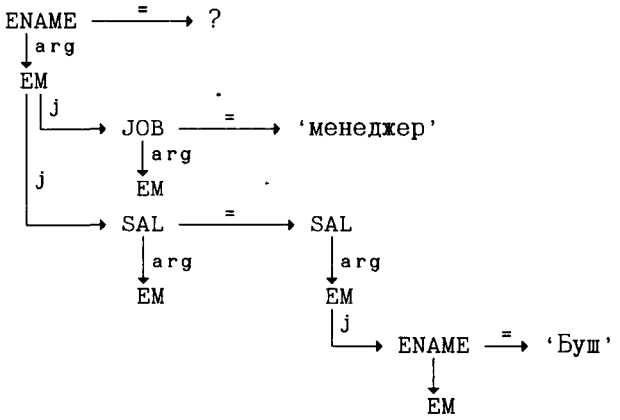
(22) *Кто имеет такую же зарплату, как Буш <как у Буша>?*

В результате нормализации этот запрос примет вид

(22') *Кто (таков, что) (этот) кто имеет зарплату, как зарплата Буша?*

Здесь сочетание *иметь зарплату, как зарплата Буша* интерпретируется как 'иметь зарплату, равную зарплате Джоунза'. Результирующая СемС запроса (22) будет выглядеть так:

(22'')



Специальное правило требуется для таких оборотов, как *3000 долларов и <или> меньше <больше>*. В отличие от обычных сочинительных оборотов типа *3000 или 4000 долларов*, подобные сочетания семантизируются не с помощью конъюнктивного или дизъюнктивного семантического отношения, а с помощью отношений '<' или '>'.

Отдельную группу составляют слова, соответствующие операторам семантического языка (COUNT 'количество', MAX 'максимум', MIN 'минимум', AVG 'среднее значение', SUM 'сумма').

Оператор COUNT вызывается словами типа ОДИН, ДВА, ТРИ,...; СКОЛЬКО, КОЛИЧЕСТВО, ЧИСЛЕННОСТЬ и т.п., которые особых трудностей для семантического анализа не представляют. Гораздо интереснее обстоит дело со словами типа МАКСИМАЛЬНЫЙ, МИНИМАЛЬНЫЙ, САМЫЙ БОЛЬШОЙ <МАЛЕНЬКИЙ, ВЫСОКИЙ, НИЗКИЙ, ...>, СРЕДНИЙ, В СРЕДНЕМ, СУММАРНЫЙ, В СУММЕ, ОБЩИЙ и т.д., которые соответствуют операторам MAX, MIN, AVG и SUM. Все эти операторы применяются не к отдельным объектам, а к множествам объектов. Сочетание *максимальная зарплата* ничего не обозначает, пока не указано, среди какого множества зарплат выбирается максимальная. Семантически более полными являются сочетания типа *максимальная зарплата в отделе 10, максимальная зарплата клерков, максимальная зарплата на фирме* и т.п. Трудность для семантического анализа этих сочетаний обусловлена тем, что выражения *отдел 10, клерки, фирма* и т.п., необходимые для интерпретации слова МАКСИМАЛЬНЫЙ и заполняющие соответствующую валентность, синтаксически с ним не связаны. Более того, возможны ситуа-

ции, когда подобные выражения в СинтС отделены от соответствующего операторного слова еще большим количеством стрелок, чем в приведенных выше примерах; ср., например, синтаксическое расстояние между словом ***максимальный*** и сочетанием ***в отделе 10*** в СинтС (23′) или в нормализованной СинтС 23′′), соответствующей запросу (23):

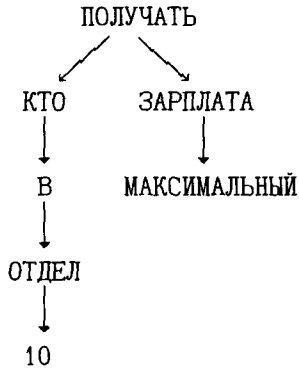
(23) ***Кто в отделе 10 получает максимальную зарплату?***

Еще больше синтаксическое расстояние между словом ***средний*** и сочетанием ***в отделе 10*** в запросе

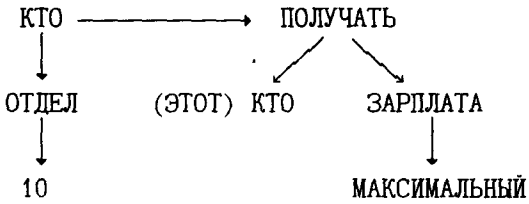
(24) ***Кто в отделе 10 получает зарплату большую, чем средняя?***;

см. ниже СинтС (24′) или нормализованную СинтС (24′′) для этого запроса.

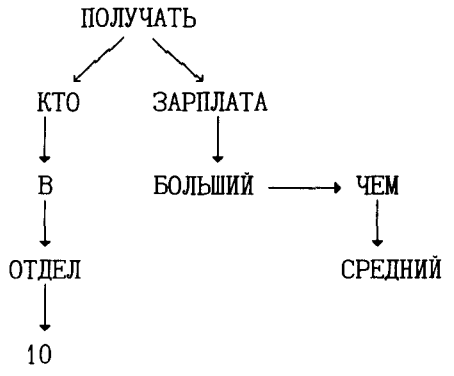
(23′)



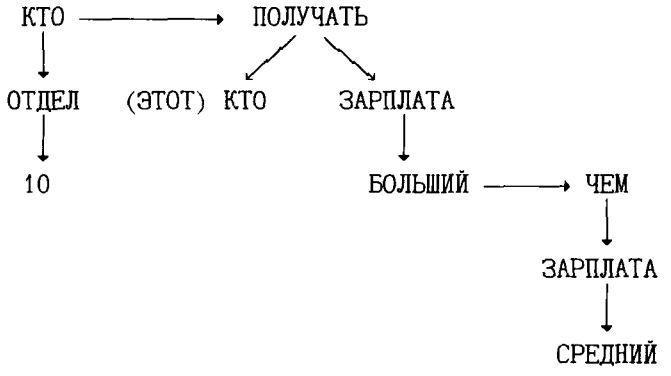
(23′′)



(24′)



(24′′)



Следует учитывать также и то обстоятельство, что отнесение выражения типа ***в отделе 10*** к операторному слову типа МАКСИМАЛЬНЫЙ не всегда однозначно. Например, запрос (24) может иметь в виду как зарплату, среднюю по отделу 10, так и зарплату, среднюю по всей фирме. Это значит, что в (24) валентность слова СРЕДНИЙ может быть заполнена сочетанием ***в отделе 10***, а может быть и ничем не заполнена, и тогда она интерпретируется в универсальном смысле. При этом разные операторные слова с разной степенью легкости могут оставлять эту валентность незаполненной. Так, в запросе

(25) ***Кто из менеджеров получает самую высокую зарплату?***

речь идет о зарплате, самой высокой среди зарплат менеджеров, а запрос

(26) ***Кто из менеджеров получает максимальную зарплату?***

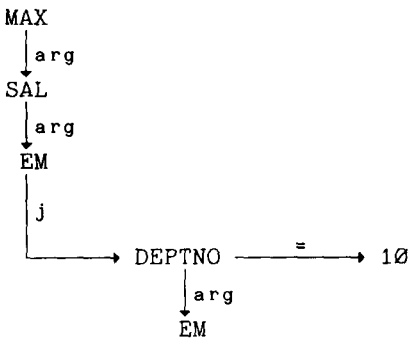
допускает два понимания. Одно из них синонимично (25), и в этом случае в ответ на запрос должна быть названа одна или несколько фамилий менеджеров. Второе понимание предполагает сравнение с зарплатой, максимальной в масштабе всей фирмы, и в этом случае уместен ответ ***"Никто"***, поскольку менеджеры

же являются самыми высокооплачиваемыми сотрудниками фирмы.

Все сказанное с очевидностью показывает, что в общем случае надежно определить интерпретацию, которую имел в виду спрашивающий, можно только в интерактивном режиме (см. об этом выше, в разд. 5.4.1).

Вопрос о заполнении валентности операторных слов типа МАКСИМАЛЬНЫЙ - не единственный подводный камень, который с ними связан. Вторая трудность состоит в том, что независимо от заполнения этой валентности сочетание типа *максимальная зарплата* по-разному встраивается в СемС запроса. Так, например, сочетание *максимальная зарплата в отделе 10* само по себе всегда значит одно и то же:

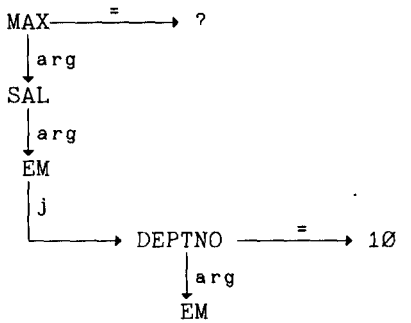
27)



Однако в разные запросы это значение входит по-разному. Например, сравнение запросов (28) и (29) с их СемС (28') и (29') показывает, что в одном случае предикаты SAL и DEPTNO входят в СемС один раз, а в другом - два, хотя внешним образом соответствующие слова фигурируют в запросе лишь однократно.

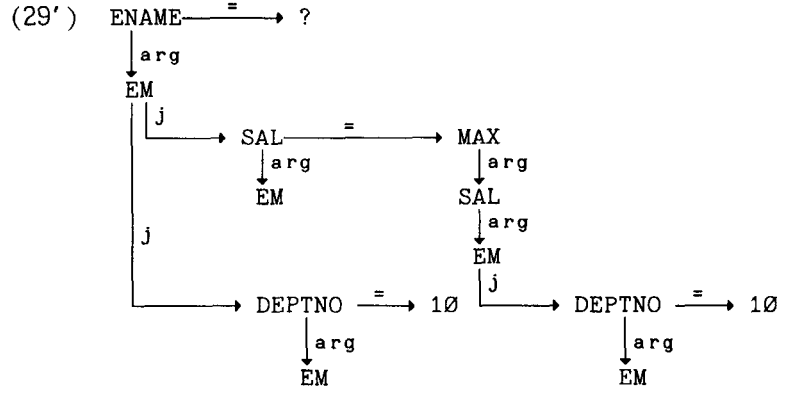
(28) *Какова максимальная зарплата в отделе 10?*

(28')



(29) *Кто получает максимальную зарплату в отделе 10?*

10 Заказ № 1538



Таким образом, внешне простая форма запроса (29) скрывает за собой содержание, которое более эксплицитно можно выразить так: 'Кто из сотрудников отдела 10 получает зарплату, равную максимальной зарплате сотрудников отдела 10?

Предложение (29) демонстрирует еще одну интересную особенность: сочетания типа ***в отделе 10*** могут выполнять ограничительную функцию не только по отношению к операторным словам типа МАКСИМАЛЬНЫЙ, но одновременно и по отношению к вопросу. В (29) это сочетание играет сразу две роли. Оно, во-первых, ограничивает множество зарплат, по которому берется максимум, а во-вторых, - множество допустимых значений вопросительной переменной ('кто из служащих отдела 10').

Теперь, когда мы рассмотрели все группы слов, имеющие аналоги в семантическом языке, можно обратиться к словам, не имеющим такого прямого соответствия. В материале, послужившем основой для разработки правил семантического анализа (свыше 200 запросов), оказалось всего несколько таких слов. Проиллюстрируем лишь два из них - наиболее показательные.

Существительные ИНТЕРВАЛ и ПРОМЕЖУТОК выступают в контекстах типа

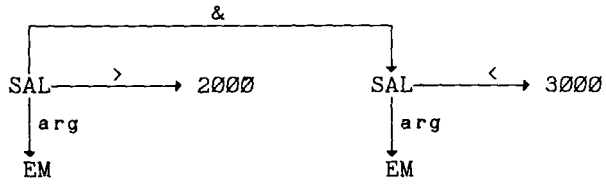
(30) ***зарплата в интервале <промежутке> между 2000 и 3000.***

Значение, передаваемое этими существительными, выражаются в семантическом языке конъюнкцией двух утверждений, задающих границы интервала[3]:

---

[3]Впрочем, семантический язык легко можно дополнить так, чтобы значение 'в интервале' передавалось более прямым способом, тем более, что язык SQL таким средством обладает.

(30′)



Такая же СемС соответствует аппроксимативной группе и в отсутствие слов типа ИНТЕРВАЛ и ПРОМЕЖУТОК *(зарплата от 2000 до 3000)*.

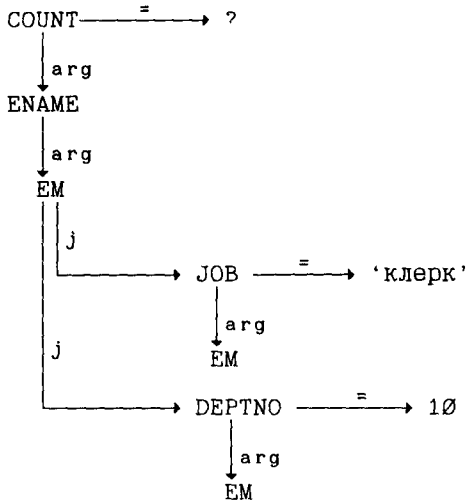
Глагол ИМЕТЬСЯ семантизируется по-разному в зависимости от контекста. Конкретнее, для семантизации существенно, каково соотношение семантических классов подлежащего и дополнения. Начнем со случая, когда подлежащее и дополнение относятся к разным классам, например:

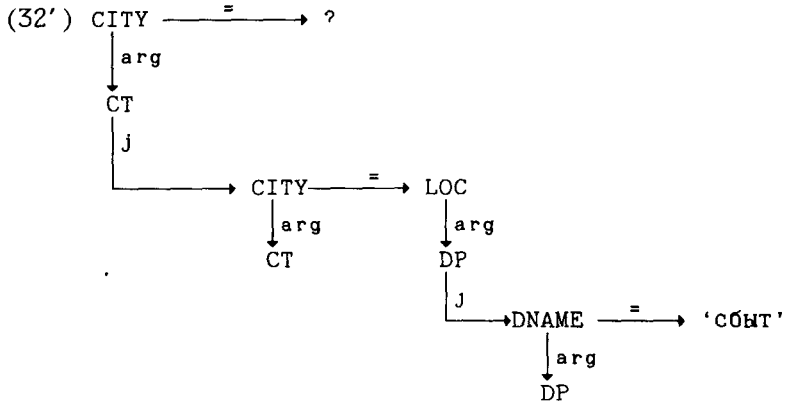
(31) ***Сколько клерков имеется в отделе 10?***

(32) ***В каком городе имеется отдел сбыта?***

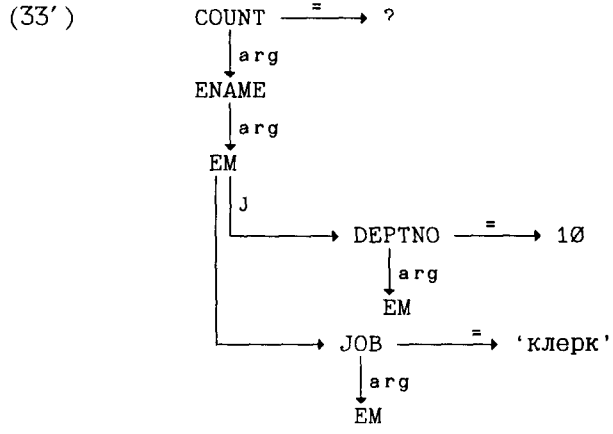
В запросе (31) подлежащее относится к классу служащих, а дополнение к классу отделов. В (32) подлежащее - из класса отделов, а дополнение - из класса городов. В контексте (31) глагол ИМЕТЬСЯ интерпретируется так же, как РАБОТАТЬ (в), и толкуется через функцию DEPTNO, а в контексте (32) - как НАХОДИТЬСЯ (в) и толкуется через функцию LOC. Поэтому СемС для этих запросов выглядят так:

(31′)





Теперь перейдем к случаю, когда подлежащее и дополнение глагола ИМЕТЬСЯ относятся к одному и тому же классу, как в (33) ***Сколько клерков имеется среди служащих отдела 10?*** СемС этого запроса имеет следующий вид:



Иначе говоря, значение запроса (33) в более эксплицитном виде можно представить так: 'Чему равно число фамилий служащих, занимающих должность клерка и работающих в отделе 10?'. Отсюда видно, что у глагола ИМЕТЬСЯ нет в этом предложении самостоятельного семантического образа и он не должен толковаться. Его роль состоит в том, чтобы сообщить, что речь идет о служащих, обладающих одновременно обоими свойствами - они являются клерками и работают в отделе 10.

Таким образом, правило семантизации глагола ИМЕТЬСЯ в запросе (33) должно произвести лишь синтаксическое преобразование: атрибутивное зависимое дополнения - ***отдела 10*** - следует перенести от дополнения к подлежащему.

Рассмотрим, наконец, случай, когда у глагола ИМЕТЬСЯ вовсе нет дополнения:

34) ***Сколько имеется служащих с окладом свыше 3000?***

После этапа нормализации этот запрос примет вид:

34') ***Сколько*** [='чему равно количество'] ***служащих с окладом свыше 3000 (таких, что) (эти) служащие имеются?***

Очевидно, что придаточное предложение со сказуемым ***имеются*** не несет здесь никакой информации и должно быть опущено.

Заметим, что два последних типа семантизации глагола ИМЕТЬСЯ состоят не в выработке каких-либо семантических элементов, а исключительно в чисто синтаксических преобразованиях: в первом случае производится перенос атрибута, а во втором - опущение семантически ненаполненных слов. Это означает, что в данных типах контекстов обсуждаемый глагол имеет, по существу, синтаксическое употребление. Это сближает его с синтаксическими отношениями, к вопросу о семантизации которых мы и переходим.

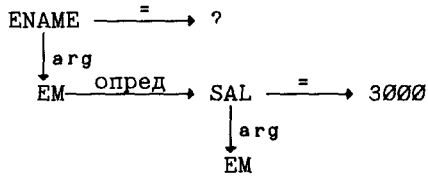
5.4.2.2. Семантизация синтаксических отношений

После работы правил семантизации лексики в узлах структуры уже практически не осталось слов естественного языка. Исключение составляют лишь чисто "синтаксические" слова вроде сочинительных союзов, отрицания или некоторых употреблений слов типа ИМЕТЬСЯ, которые естественно семантизировать вместе с синтаксическими отношениями. В ходе лексических преобразований "заодно" исчезли и многие синтаксические отношения. Например, запрос типа

(35) ***Укажите служащих, получающих зарплату 3000 долларов.***

превратился к этому моменту в почти готовую СемС

(35')



Единственное, что еще остается здесь сделать, это превратить определительное синтаксическое отношение в семантическое отношение 'J' Помимо подобных тривиальных переименований, правила семантизации синтаксических отношений решают и ряд менее тривиальных задач, связанных в основном с

обстоятельственными, атрибутивными, элективными и сочинительными конструкциями.

Обстоятельственные конструкции

В целом обстоятельственные конструкции для нашей предметной области нетипичны. В ней нет ни причин, ни времени, ни способов действия, ни условий, при которых оно может совершаться, ни даже самих действий. Достаточно распространенными являются лишь локативные и детерминирующие обстоятельства, которые мы сейчас и рассмотрим. (Реальны также и обстоятельства, выраженные деепричастными оборотами, но они исчезают уже на этапе нормализации - см. разд. 5.4.1.)

Локативные обстоятельства, возможные в запросах по нашей предметной области, можно проиллюстрировать следующими тремя примерами.

(36) *В каком отделе все клерки получают больше 3000?*

(37) *В каком штате отдел возглавляет Кларк?*

(38) *В каком отделе количество служащих равно десяти?*

Подобные обстоятельства, как выяснилось, во многом отличаются от типичных локативных обстоятельств или детерминантов типа

(39) *В нашем городе Кларк выступил с концертом.*

(40) *В каком штате Кларк заработал 5000 долларов?*

Подробное сопоставление обстоятельств этих двух типов дается в работе [Богуславский, 1991]. Здесь мы ограничимся лишь несколькими замечаниями.

Обстоятельства, выступающие в предложениях (39)-(40), - их можно назвать собственно локативными обстоятельствами - задают чисто пространственные координаты. Они фиксируют точку или область физического пространства, в которой локализуется некоторая ситуация и/или ее участники. В каноническом случае речь идет не только об участниках ситуации, но и о ситуации в целом. В предложении (39), например, говорится, что не только Кларк находился в нашем городе, но и что сама ситуация выступления имела место именно там. Ситуации, которые нельзя мыслить как разворачивающиеся в физическом пространстве, не допускают подобных обстоятельств. Таковы, например, постоянные свойства, отношения, так называемые устойчивые состояния и др. Предикаты, обозначающие такие ситуации, включают слова типа БЕСПОКОИТЬ, ВЕРИТЬ,

ЕЛАТЬ, ЗНАТЬ, НЕНАВИДЕТЬ и многие другие [Рахилина, 1988, : 92-93]: *верить в городе, *знать в транспорте, *ненавидеть в саду.

Другой тип локативных обстоятельств, к которому относятся предложения (36)-(38) и к которому мы хотели бы привлечь особое внимание, мы будем условно называть (локативно-) миропорождающими обстоятельствами. Различие между собственно локативными и миропорождающими обстоятельствами удобно проиллюстрировать примером, который допускает обе интерпретации:

(41) *Здесь солнце не заходит.*

В первой интерпретации - собственно локативной - слово ЗДЕСЬ фиксирует ту часть небосвода, в которой заходящее солнце уходит (или не уходит) за горизонт. В этой интерпретации предложение (41) может быть продолжено, например, так: *Здесь солнце не заходит, оно заходит левее, за лесом.*

При второй интерпретации речь идет не о местонахождении солнца в момент захода, а о мире, в котором имеет или не имеет место ситуация захода солнца. Возможное продолжение (41) в этом случае может быть, например, таким: *Здесь солнце не заходит - мы ведь за Полярным кругом.*

Одно из наиболее ярких отличий миропорождающих обстоятельств от собственно локативных состоит в том, что на них не распространяются те ограничения на сочетаемость с глаголами, о которых мы говорили выше:

(42) *Он ненавидит жадность в саду.

(43) *В нашем городе все ненавидят жару.*

(44) *Он знает математику в автобусе.

(45) *Здесь знают немецкий язык.*

В силу самого характера миропорождающего значения естественно предположить, что сочетаемость этих обстоятельств с глаголами вряд ли может быть чем-то ограничена.

Соотношение между собственно локативным и миропорождающим значением в значительной мере определяется различием между понятием физического пространства и более общим понятием мира, которое включает физическое пространство в качестве одного из компонентов. Семантика миропорождающих обстоятельств двойственна, и это не может не сказываться на том, к каким пропозициям они могут применяться.

С одной стороны, в значении этих обстоятельств сильна

локативная составляющая. Пропозиция, которую определяет обстоятельство, должна мыслиться в рамках данного мира. Поэтому невозможно присоединение такого обстоятельства к пропозиции, которая локализуется в пространстве, не входящем в состав данного мира; ср. пару (46)-(47):

(46) *В нашем городе школьники по субботам не учатся.*

(47) **В нашем городе школьники провели лето в Крыму.*

Предложение (47) неправильно, поскольку ситуация "проведение лета" локализована в Крыму, что несовместимо с ее локализацией в мире нашего города.

С другой стороны, поскольку понятие мира шире, чем понятие физического пространства, миропорождающие обстоятельства могут присоединяться к абстрактным суждениям, для которых нормальная пространственная локализация не имеет смысла:

(48) *В Индии корова - священное животное.*

(49) *В нашем городе дважды два не всегда равно четырем.*

Одна из самых важных особенностей миропорождающих обстоятельств касается их соотношения с именными группами, входящими в состав предложения. Интересующий нас здесь аспект этого соотношения состоит в том, что такие обстоятельства способны ограничивать экстенсионал имен, входящих в их сферу действия.

Так, если во фразе

(50) *Средняя зарплата клерков составляет 3000 долларов*

экстенсионал имени *клерки* ничем не ограничен, то во фразе

(51) *В отделе 10 <в Чикаго> средняя зарплата клерков составляет 3000 долларов*

он ограничивается отделом 10 <городом Чикаго>.

Такую же функцию ограничения экстенсионала выполняют и атрибуты имени. Именно поэтому обстоятельство в подобных предложениях можно превратить в атрибут: предложения (36)-(38) синонимичны предложениям (38')-(40').

(38') *Все клерки какого отдела получают больше 3000?*

(39') *Отдел какого штата возглавляет Кларк?*

(40') *Количество служащих какого отдела равно десяти?*

Отличие миропорождающего обстоятельства от атрибута касается не столько значения, сколько объема сферы действия: атрибут воздействует только на то имя, которому он подчинен синтаксически, в то время как обстоятельство может харак-

теризовать все имена, входящие в его сферу действия:

52) *В отделе 10 максимальная зарплата клерков равна минимальной зарплате аналитиков* - 'максимальная зарплата клерков отдела 10 равна минимальной зарплате аналитиков этого же отдела'.

Указанное соотношение обстоятельств и атрибутов имеет принципиальное значение для семантизации обстоятельств, поскольку оно позволяет сводить их к атрибутам. Конкретнее, правило семантизации подобных обстоятельств превращает их в атрибуты всех тех имен, которые допускают подобный атрибут, но не имеют его. Например, в предложении (52) атрибут появляется у существительных *клерки* и *аналитики*, а в предложении

(53) *В отделе 10 максимальная зарплата клерков равна минимальной зарплате аналитиков отдела 20*

- только у первого из них, поскольку существительное *аналитики* уже охарактеризовано с точки зрения принадлежности к отделу. По этой же причине в обоих предложениях не приобретает атрибута и существительное *зарплата*. Хотя в принципе оно допускает атрибут *отдела 10* со значением 'служащих отдела 10' (ср. *средняя зарплата отдела 10*), но в предложениях (52) и (53) эта его способность уже реализована в сочетаниях *зарплата клерков* и *зарплата аналитиков*.

Похожее правило обслуживает и детерминирующие обстоятельства со словом *каждый* в запросах типа:

(54) *Для каждого служащего указать зарплату ⇒ Указать зарплату каждого служащего;*

(55) *Для каждой должности указать количество служащих ⇒ Указать количество служащих (по) каждой должности.*

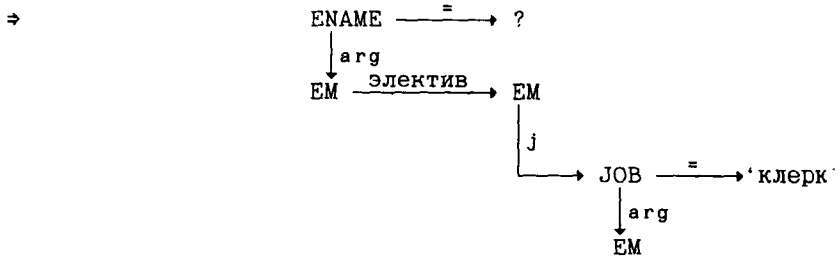
Атрибутивные и элективные конструкции

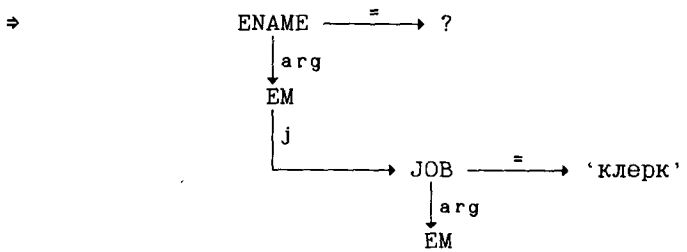
Конструкции этого типа, поступающие на семантизацию, подразделяются на два класса: конструкции, элементами которых являются переменные (тождественные или нетождественные), и конструкции, состоящие из переменной и функции.

Приведем примеры на оба эти класса. Соответствующие правила семантизации будут очевидны из примеров.

1) Конструкции из тождественных переменных:

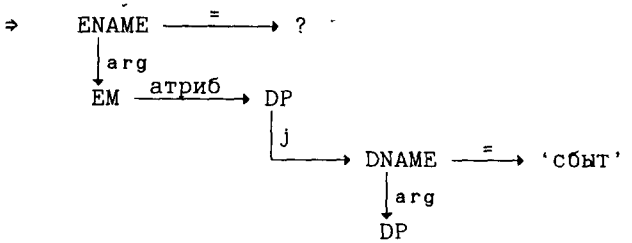
(56) *Кто из клерков...*

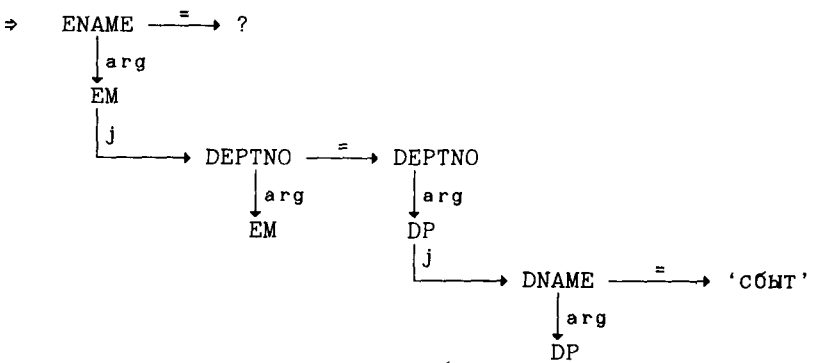




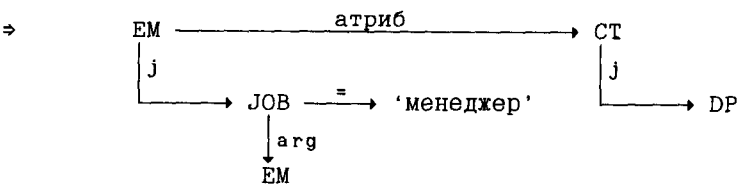
2) Конструкции из нетождественных переменных:

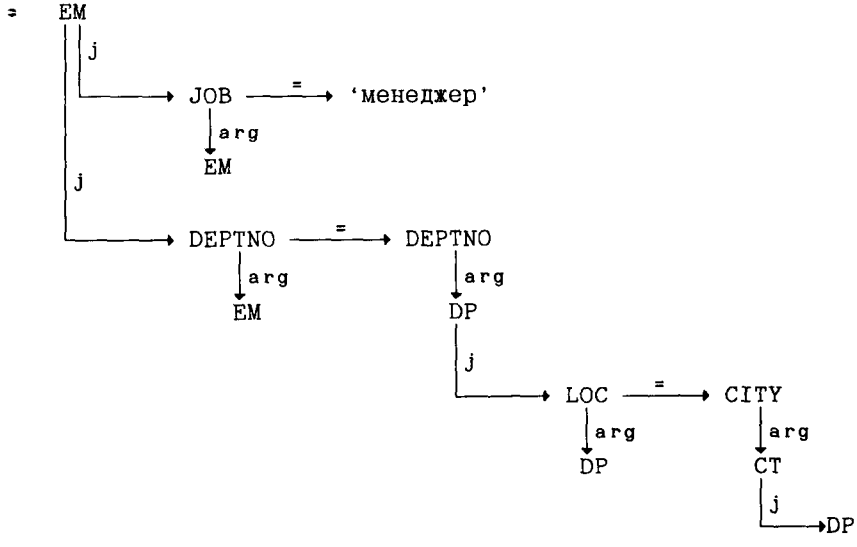
(57) ***Кто в отделе сбыта...***





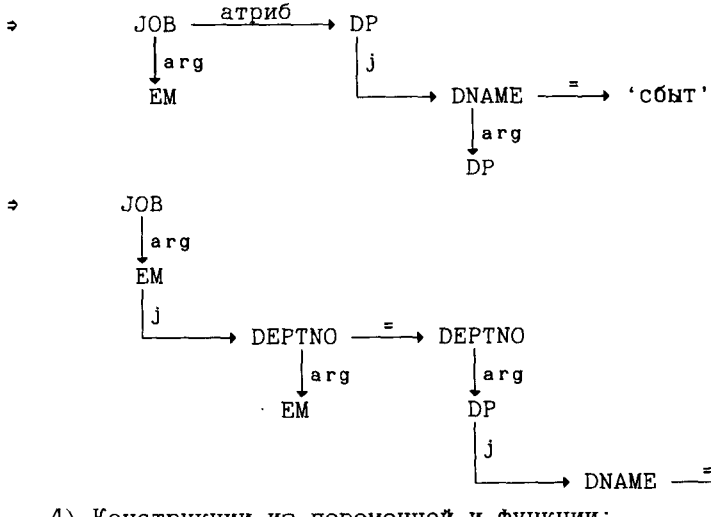
(58) ***менеджер из города, который***





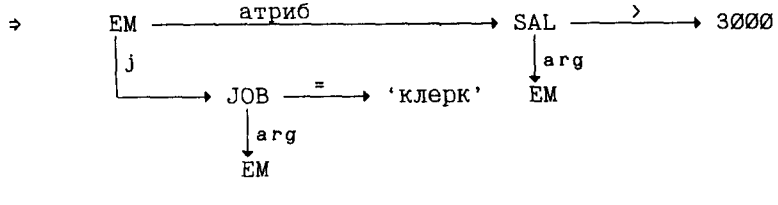
3) Конструкции из функции и переменной:

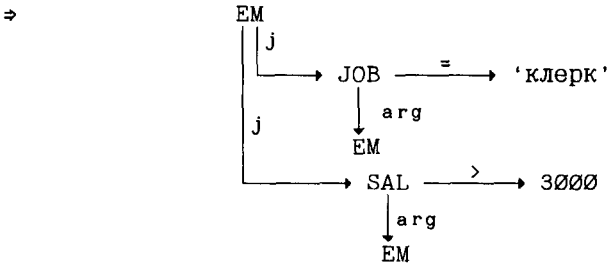
59) ***должности в отделе сбыта***



4) Конструкции из переменной и функции:

(60) ***клерки с зарплатой свыше 3000***





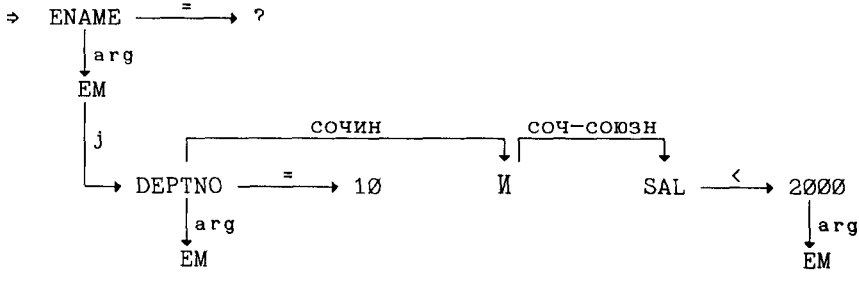
Сочинительные конструкции

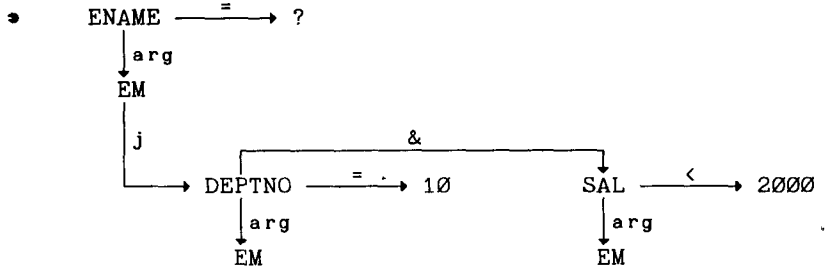
Как следует из определения СемС, которое было дано в разд. 5.3.4, информация, заключенная в сочинительных конструкциях естественного языка, передается в СемС двумя семантическими отношениями - конъюнктивным (&) и дизъюнктивным (v). Эти отношения могут связывать в СемС только функции и операторы. Поэтому сочинительные конструкции, построенные из элементов других типов (например, такие конструкции, как *клерки и менеджеры; Джоунз или Смит* и т.п.), должны в конечном счете быть развернуты в более эксплицитные структуры.

В ряде ситуаций сочинительное развертывание производится уже на этапе нормализации (см. разд. 5.4.1). Это происходит тогда, когда сочинительная конструкция находится в контексте сравнительной и они должны быть развернуты совместно. В остальных случаях эта работа осуществляется на этапе семантизации. Конкретнее, правила семантизации сочинительных конструкций обеспечивают выполнение следующих задач.

Прежде всего, сами сочинительные союзы, наряду с сочинительными и сочинительно-союзными синтаксическими отношениями, должны быть переработаны в соответствующие семантические отношения, например:

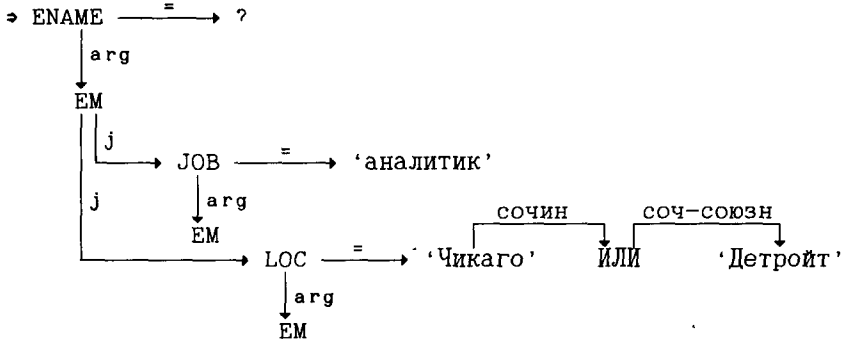
(61) *Кто работает в отделе 10 и зарабатывает меньше 2000?*





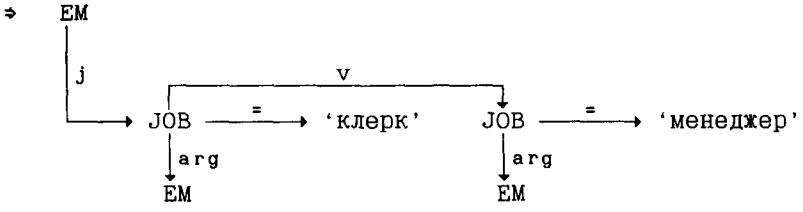
В этом примере союз И в момент семантизации связывает функциональные элементы (DEPTNO и SAL), которые имеют полное право быть соединены семантическим отношением *&*. Если же сочинительную конструкцию образуют элементы других типов, то, как уже говорилось, необходимо провести сочинительное развертывание:

(62) ***Какие аналитики работают в Чикаго или Детройте?***

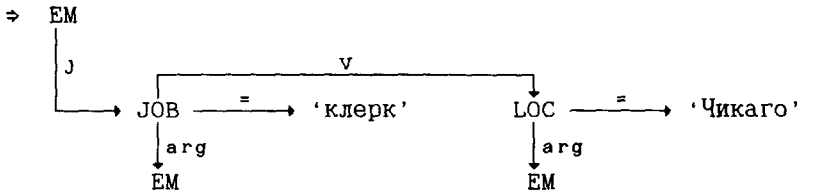


Наконец, необходимо учитывать, что в некоторых контекстах союз И должен переводиться не конъюнктивным, а дизъюнктивным семантическим отношением. Такую функцию союз И (в нашем мире) имеет в контексте нефункциональных существительных (т.е. тогда, когда он соединяет не слова типа ЗАРПЛАТА, а слова типа КЛЕРК). Например:

(63) ***клерки и менеджеры*** (='служащие, являющиеся клерками или менеджерами')

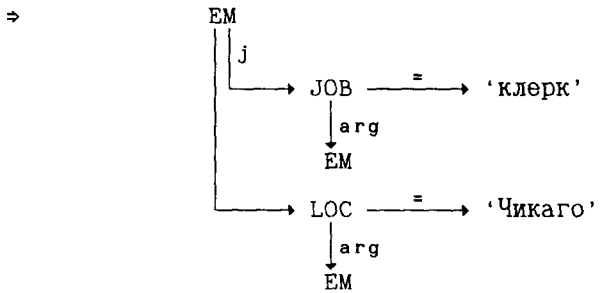


(64) ***клерки и жители Чикаго*** (='служащие, являющиеся клерками или живущие в Чикаго')



Чтобы преодолеть действие этой закономерности и навязать конъюнктивное понимание союза, следует использовать специальные средства, например, наречие ОДНОВРЕМЕННО:

(65) ***клерки и одновременно жители Чикаго*** (='служащие, являющиеся клерками и живущие в Чикаго')



5.4.3. Канонизация СемС

Правила нормализации и семантизации, о которых говорилось в предыдущих разделах, в основном, проделали всю содержательную работу по преобразованию СинтС в СемС. Однако полученный результат еще не является полноценной СемС. Это обусловлено двумя обстоятельствами.

Во-первых, правила семантизации носят локальный характер. Это значит, что каждое правило вырабатывает фрагмент СемС, соответствующий отдельной лексической единице или синтаксической конструкции естественного языка. Структура, получающаяся после этих преобразований, может обладать тем недостатком, что в ней недостаточно сглажены стыки между отдельными фрагментами. В этом случае необходимо провести специальные преобразования, осуществляющие оптимальную подгонку отдельных фрагментов СемС друг к другу. В их число входят такие операции, как устранение разного рода тавтологий, стягивание кореферентных функций, освобождение констант от лишних слуг, подъем предикатных семантических от-

ношений (типа '=', '≠', '<', '>' и т.п.) от функций к операторам, вытягивание конъюнктивно-дизъюнктивных цепочек, подъем аргумента по конъюнктивно-дизъюнктивной цепочке и некоторые другие преобразования. После проведения этих преобразований СемС становится вполне прозрачной и связной.

Во-вторых, даже после этого СемС сохраняет ряд черт, которые, не затрудняя ее восприятие человеком, отличают ее от канонической СемС, заданной формальным определением в разд. 5.3.4. Сюда относятся главным образом случаи разного рода рассогласований между функциями и их аргументами. Специальные правила снимают рассогласование, устанавливая своего рода мостики между рассогласованными элементами.

Правила обоих типов носят довольно технический характер, и вряд ли целесообразно останавливаться на них подробно. Мы проиллюстрируем по одному правилу из каждой группы, чтобы дать представление о характере решаемых задач.

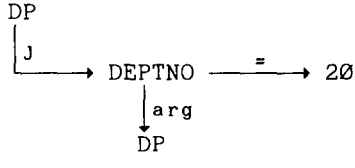
Устранение тавтологии

Рассмотрим сочетание ***отдел 20*** в составе запроса

66) ***Где расположен отдел 20?***

СемС этого запроса получается с использованием правила

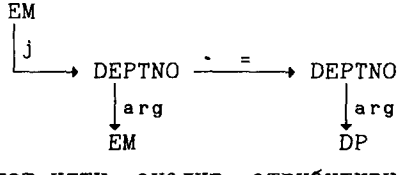
67) ***отдел 20*** ⇒



'отдел, номер которого - 20'.

С другой стороны, имеется общее правило семантизации атрибутивных и комплетивных конструкций, в которых главным членом является EM, а зависимым - DP, обслуживающее большой класс сочетаний типа ***клерки отдела сбыта, менеджер из бухгалтерии, служащие отделов, расположенных в Чикаго*** и т.п. Это правило, как мы показали выше, производит следующее преобразование:

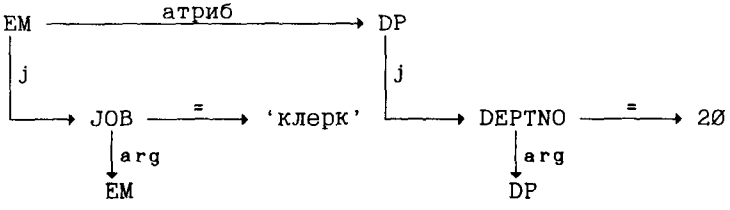
(68) EM —атриб→ DP ⇒



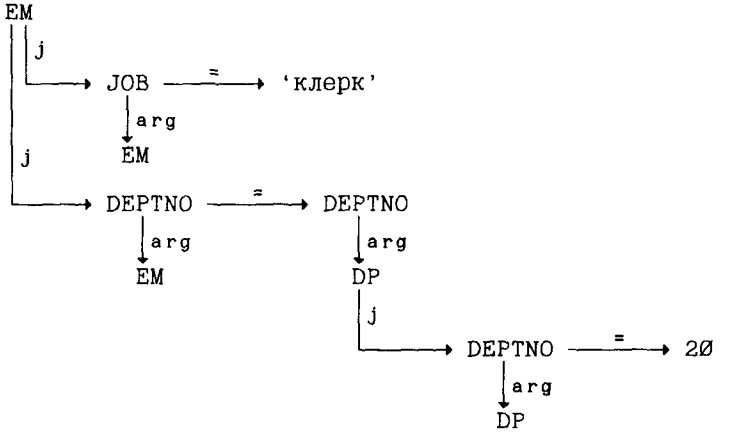
Рассмотрим теперь как будет идти анализ атрибутивного сочетания

(69) *клерки из отдела 20.*

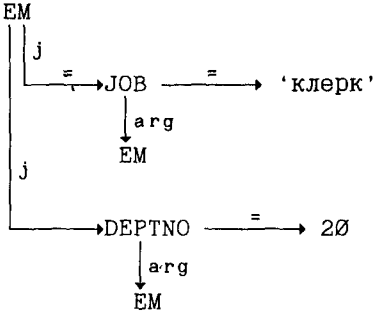
Правило нормализации опустит предлог ИЗ. С помощью толкования слова КЛЕРК и правила (67) получится структура

(70)


К этой структуре применимо правило (68), которое даст в результате

(71)


Буквальное прочтение этой структуры выглядит так: 'служащий, имеющий должность клерка и работающий в отделе, номер которого равен номеру отдела, номер которого равен 20'. Налицо тавтология. Более компактно нужный смысл можно передать так:

(72)


Для того, чтобы не иметь тавтологичных структур типа 71), можно пойти по одному из двух путей.

1) Вместо одного правила (67) семантизации сочетаний типа *отдел 20*, применяющегося во всех контекстах, иметь два менее общих правила, которые будут давать в контекстах типа 66) один результат, а в контекстах типа (69) - другой, не содержащий тавтологии.

2) Иметь сформулированные выше общие правила, а возникающую из-за их общности тавтологию устранять единым правилом на уже готовой СемС.

Поскольку ситуация, подобная описанной, возникает для сочетаний многих типов на разных участках семантизации, преимущество второго пути представляется бесспорным. Только он позволяет получить достаточно компактное и общее описание и не приводит к чрезмерному дроблению правил.

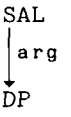
Устранение рассогласования

В запросах типа

(73) ***Какова средняя зарплата отдела сбыта?***

выступает сочетание ***зарплата отдела***, которое порождает фрагмент СемС

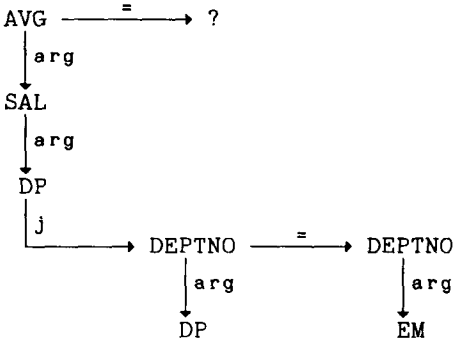
(73′)



Этот фрагмент, будучи вполне понятным, формально некорректен, поскольку функция SAL, по определению, не допускает в качестве аргумента переменной DP. Это требование, как и некоторые другие, заложенные в определение СемС, вызвано стремлением обеспечить определенный изоморфизм между СемС и выражениями языка SQL.

С точки зрения SQL аргументная связь между функцией и переменной в СемС интерпретируется как информация о том, в какой таблице БД следует искать значения атрибута, соответствующего данной функции. Поэтому аргументная связь между SAL и DP для SQL неприемлема: атрибут SAL принадлежит не таблице DP, а таблице EM. Для снятия подобных несоответствий в SQL существует механизм сцепления таблиц. На языке СемС это означает введение в нее предикации, отражающей связь между разными переменными. С помощью такого преобразования СемС запроса (73) приобретет следующий вид:

(74)



После этапа канонизации СемС полностью готова к преобразованию в SQL-выражение, предназначенное для ввода в СУБД Подобное преобразование не представляет никаких трудностей, потому что, как уже отмечалось, в определение СемС изначально заложен изоморфизм с SQL-выражениями. Поскольку результатом этого преобразования является недревесный объект, его нецелесообразно описывать на том языке (см. гл. 2), на котором записывается основной массив правил ЛП. Оно осуществляется специальной процедурой, которая помимо СемС существенным образом использует информацию о кореферентности.

На этом мы закончим изложение лингвистических средств перехода от СинтС запроса к его СемС. В следующем разделе мы приведем образцы правил нормализации, семантизации и канонизации на формальном языке, а затем обратимся к алгоритмическому обеспечению этого перехода.

5.4.4. Образцы правил семантического анализа

5.4.4.1. Общее правило нормализации

Формальная запись правила

```
REG:NORMAL1.52     ПРЕОБРАЗОВАНИЕ СТРАДАТЕЛЬНОГО ПРИЧАСТНОГО
CHECK              ОБОРОТА В ПРИДАТОЧНОЕ ПРЕДЛОЖЕНИЕ
1.1 =(X,СТРАД,ПРИЧ)&DEP(X,Z,ОПРЕД)
N:Ø1               У ПРИЧАСТИЯ ЕСТЬ АГЕНТИВНОЕ ДОПОЛНЕНИЕ
CHECK
1.1 DOM(X,Z1,АГЕНТ)
DO
1 ŻAMUZHAR X(ЛИЧ)
2 DOBUZHAR:X(ИЗЪЯВ)
3 STERUZHAR.X(СТРАД)
4 DOBRUZ:W(ФИКТ-ЛЕКС)
```

```
5 PERLEKRUZ:Z(W)
6 PERUZINOM:W(X)
7 SVUZOT:(X,W,1-КОМПЛ)
8 IZOT:(X,Z,ОПРЕД)-(X,Z,РЕЛЯТ)
9 IZOT:(X,Z1,АГЕНТ)-(X,Z1,ПРЕДИК)
10 SVUZREF:(Z,W,*)
N:02                 У ПРИЧАСТИЯ НЕТ АГЕНТИВНОГО ДОПОЛНЕНИЯ
CHECK
DO
1 ZAMUZHAR:X(ЛИЧ)
2 DOBUZHAR:X(ИЗЪЯВ,МН,3-Л)
3 STERUZHAR:X(RGNR,RAN,RNMB,RCS,СТРАД)
4 DOBRUZ:W(ФИКТ-ЛЕКС)
5 PERLEKRUZ:Z(W)
6 PERUZMINOM.W(X)
7 SVUZOT:(X,W,1-КОМПЛ)
8 IZOT:(X,Z,ОПРЕД)-(X,Z,РЕЛЯТ)
9 SVUZREF.(Z,W,*)
```

Комментарий к правилу

Правило предназначено для превращения причастных оборотов, содержащих страдательное причастие, в личное предложение. При этом отсутствующий актант причастия восстанавливается и связывается кореферентным отношением с определяемым существительным: ***отдел, возглавляемый Смитом ⇒ отдел, такой, что (этот) отдел возглавляет Смит.*** Заметим, что обработка страдательных причастий очень близка, но не вполне тождественна тому, что нужно делать с причастиями действительного залога: в случае действительных причастий восстановленный актант становится подлежащим личного глагола, а в случае страдательных причастий - его первым дополнением.

Правило содержит два лодправила. Первое обрабатывает конструкции с агентивным дополнением (***отдел, возглавляемый Смитом***), а второе - конструкции, в которых такого дополнения нет (***отдел, называемый "бухгалтерия"***).

5.4.4.2. Трафаретное правило нормализации

Формальная запись правила

```
REG:NORMAL.12        УСТРАНЕНИЕ УКАЗАТЕЛЬНЫХ МЕСТОИМЕНИЙ
TAKE.1               С УСТАНОВЛЕНИЕМ КОРЕФЕРЕНТНОСТИ
LOC:R
```

```
R:1-КОМПЛ/ОБСТ
N:Ø1
CHECK
1.1 DEP-EQU(X,Z1,ОПРЕД)
1.2 IDOM-EQU(Z1,Z2,*,S)/IDEP(X,W,*)&IDOM-EQU(W,Z2,*,S)
1.3 ORD(Z2,X)
1.4 COLEX(Z1,Z2)/CODES(Z1,Z2)
DO
1 SVUZREF:(Z2,Z1,*)
2 STERUZ:X
```

Комментарий к правилу

Задача этого правила - обнаружить антецедент указательного местоимения (типа *этот, такой, данный, указанный*) и установить кореферентное отношение между этим антецедентом и словом, к которому относится местоимение. В зоне CHECK перечислены условия, при которых узел Z2 может считаться антецедентом местоимения X. Во-первых, узел Z2 должен располагаться левее X-а, а во-вторых, он должен находиться в определенном соотношении с существительным Z1, с которым синтаксически связано местоимение X: лексемы в узлах Z1 и Z2 должны либо быть тождественны (*...фамилии клерков* [Z2] *и зарплата этих клерков* [Z1]), либо иметь общие семантические признаки (*...фамилии клерков* [Z2] *и зарплата этих служащих* [Z1]).

5.4.4.3. Словарное правило нормализации

Формальная запись правила

```
REG:NORMAL1.13  РАЗВОРАЧИВАНИЕ КОНСТРУКЦИЙ ТИПА ТАМ ЖЕ, ГДЕ
LOC:RR1,RR2,ALFA,BETA
RR1:1-КОМПЛ/2-КОМПЛ/ПРЕДИК/ОБСТ
RR2.1-КОМПЛ/2-КОМПЛ/ПРЕДИК/ОБСТ
ALFA:S/A/ADV/NUM
BETA:НЕПРОШ/ПРОШ/БУД
N:Ø1
CHECK
1.1 DOM-EQU(X,Z1,СРАВН-СОЮЗН,ALFA)&DEP-LEXR(X,Z2,СООТНОС,
     LR)&DEP-EQU(Z2,Z3,RR1,V)&=(Z3,BETA)& DOM(Z3,Z4,RR2)&
     DOM-LEXR(Z2,Z5,ОГРАНИЧ,ЖЕ)
DO
1 DOBRUZ:W(ФИКТ-ЛЕКС)
```

```
2 PERLEKRUZ:Z3(W)
3 PERUZSLEDNOM W(X)
4 SVUZOT:(X,W,СРАВН-СОЮЗН)
5 PERUZOT:(X,Z1,*)-(W,Z1,RR2)
6 STERUZ:Z5
```

Комментарий к правилу

Правило записывается в словарной статье слова ГДЕ. Оно применяется к сравнительным конструкциям типа ***работает там же, где Смит*** и восстанавливает в них глагол, попутно устраняя избыточный элемент ЖЕ: ***работает там, где работает Смит.***

5.4.4.4. Общее правило семантизации

Формальная запись правила

```
REG:SEMANT2 55  БЕССОЮЗНАЯ СОЧИНИТЕЛЬНАЯ КОНСТРУКЦИЯ
LOC:R
R:EQU/NON-EQU/MOR/LES/MOR-EQU/LES-EQU
N:Ø1            КОНЪЮНКТИВНОЕ СОЧИНЕНИЕ ФУНКЦИЙ
CHECK
1.1 ^#(X,FNC,OPR)&DOM-EQUN(X,Z,СОЧИН,FNC,OPR)
DO
1 IZOT:(X,Z,*)-(X,Z,CON)
N:Ø2            КОНЪЮНКТИВНОЕ СОЧИНЕНИЕ CONST/VAR И ФУНКЦИИ
CHECK
1.1 ^#(X,CONST,VAR)&DEP-EQUN(X,Z,R,FNC,OPR)&DOM-EQUN(X,Z1,
    СОЧИН,FNC,OPR)&^DOM-LEXR(Z1,*,СОЧИН,ИЛИ)
DO
1 PERUZOT.(X,Z1,*)-(Z,Z1,CON)
N:Ø3            ДИЗЪЮНКТИВНОЕ СОЧИНЕНИЕ ФУНКЦИЙ
CHECK
1.1 ^#(X,FNC,OPR)&DOM-EQUN(X,Z,СОЧИН,FNC,OPR)&
    DOM-LEXR(Z,*,СОЧИН,ИЛИ)
DO
1 IZOT:(X,Z,*)-(X,Z,DIS)
N:Ø4            ДИЗЪЮНКТИВНОЕ СОЧИНЕНИЕ CONST/VAR И ФУНКЦИИ
CHECK
1.1 ^#(X,CONST,VAR)&DEP-EQUN(X,Z,R,FNC,OPR)&DOM-EQUN(X,Z1,
    СОЧИН,FNC,OPR)&DOM-LEXR(Z1,*,СОЧИН,ИЛИ)
DO
1 PERUZOT:(X,Z1,*)-(Z,Z1,DIS)
```

Комментарий к правилу

В этом правиле проводится обработка бессоюзных сочинительных конструкций, в которых участвуют функции. Правило применяется тогда, когда все полнозначные слова уже семантизированы.

Прежде всего, различаются ситуации конъюнктивного сочинения (подправила 1 и 2) и дизъюнктивного сочинения (подправила 3 и 4). В первом случае вырабатывается семантическое отношение CON (&), а во втором - DIS (v). Кроме того, необходимо учитывать, что первым членом сочинительной конструкции может быть не только функция (как в случае ***указать фамилии клерков, их адреса...***), но и переменная или константа (как в случае ***указать клерков, их адреса...***). Ситуации первого типа обслуживаются подправилами 1 и 3, а второго типа - подправилами 2 и 4.

5.4.4.5. Трафаретное правило семантизации

Формальная запись правила

```
REG SEMANT1.13      СЕМАНТИЗАЦИЯ СЛОВ, ОБОЗНАЧАЮЩИХ ДОЛЖНОСТЬ
TAKE:1
N:Ø1                СОЧЕТАНИЯ ТИПА ЗАРПЛАТА МЕНЕДЖЕРА
CHECK
1.1 ^COREF(*,X)&^DOM(X,*,ПРЕДИК)&^DEP-EQU(X,*,EQU,FNC,
     'ДОЛЖН')
DO
1 ZAMAUZ:X(EM)
2 DOBAUZ:Z(JOB)
3 SVUZOT:(X,Z,JOT)
4 PERUZSLEDNOM:Z(X)
5 DOBAUZ:Z2(EM)
6 SVUZREF:(X,Z2,*)
7 SVUZOT:(X,Z2,ARG)
8 PERUZSLEDNOM:Z2(Z)
9 DOBAUZ:Z1(LA)
1Ø SVUZOT:(Z,Z1,EQU)
11 PERUZSLENOM:Z1(Z2)
N:Ø2                СОЧЕТАНИЯ ТИПА РАБОТАЕТ МЕНЕДЖЕРОМ
CHECK
1.1 DEP-LEXA(X,*,EQU,JOB)
DO
```

```
: ZAMAUZ:LA
N:Ø3                СОЧЕТАНИЯ ТИПА МЕНЕДЖЕР, КОТОРЫЙ
CHECK
:.1 COREF(*,X)
DO
: ZAMAUZ:X(EM)
N:Ø4                СОЧЕТАНИЯ ТИПА КТО МЕНЕДЖЕР
CHECK
:.1 DOM(X,Z,ПРЕДИК)
DO
1 ZAMAUZ:X(JOB)
2 DOBAUZ:W(LA)
3 SVUZOT:(X,W,EQU)
4 PERUZSLEDNOM:W(Z)
5 IZOT:(X,Z,*)-(X,Z,ARG)
```

Комментарий к правилу

Данное правило производит семантизацию слов типа КЛЕРК, МЕНЕДЖЕР, АНАЛИТИК, ПРЕЗИДЕНТ и т. п. в различных контекстах. Правило содержит параметр LA, который показывает, какую константу необходимо ввести в структуру. Например, в словарной статье слова КЛЕРК в качестве значения параметра LA дается константа 'клерк'.

Различаются четыре типа контекстов, в которых слова типа КЛЕРК семантизируются по-разному (см. разд. 5.4.2.1). Они описываются в четырех подправилах. Первое подправило обрабатывет наиболее общий контекст, в котором слово типа КЛЕРК толкуется как 'служащий, должность которого есть клерк'. Остальные контексты более специфичны. Во втором подправиле учитываются ситуации, когда слово типа КЛЕРК связано с функцией JOB семантическим отношением EQU (≈). Такое происходит, когда слово выступает в сочетаниях типа *работать клерком*. В этом случае семантическим коррелятом является просто соответствующая константа. Третье подправило применяется тогда, когда обрабатываемое слово X имеет антецедент, с которым оно связано отношением кореферентности. Это отношение возникает в ходе нормализации из местоимения, например: *клерки, которые живут в Чикаго ⇒ клерки, такие, что (эти) клерки* [X] *живут в Чикаго*. В подобных контекстах X просто заменяется на переменную EM. Наконец, последнее подправило

предназначено для обработки предложений, в которых X является именной частью сказуемого: ***Кто менеджер в отделе 10?***

5.4.4.6. Словарное правило семантизации

Формальная запись правила

```
REG:SEMANT1.23     СЕМАНТИЗАЦИЯ СЛОВА РАБОТАТЬ:
TAKE:1             СОЧЕТАНИЯ ТИПА РАБОТАТЬ КЛЕРКОМ
LOC:R
R:2-КОМПЛ/ОБСТ
CHECK
1.1 DOM-EQU(X,Z,ПРЕДИК,'ЧЕЛОВЕК')&DOM(X,Z1,1-КОМПЛ)
N:Ø1
CHECK
1.1 ^DOM(X,*,R)
DO
1 ZAMAUZ:X(JOB)
2 IZOT:(X,Z1,*)-(X,Z1,EQU)
N:Ø2
CHECK
DO
1 DOBAUZ:W(JOB)
2 SVUZOT:(Z,W,JOT)
3 PERUZMANOM:W(Z)
4 DOBAUZ:W1(EM)
5 SVUZOT:(W,W1,ARG)
6 SVUZREF:(Z,W1,*)
7 PERUZSLEDNOM:W1(W)
8 PERUZOT:(X,Z1,*)-(W,Z1,EQU)
9 PERGRSLEDNOM:Z1(W1)
```

Комментарий к правилу

Задача правила состоит в том, чтобы выработать на основе сочетания типа ***работать клерком*** семантическое поддерево

$$JOB \xrightarrow{=} \text{'клерк'}$$

Осложняет дело то обстоятельство, что для глагола РАБОТАТЬ существуют и другие возможности семантизации: например, сочетания типа *работать в бухгалтерии* порождают функцию DEPTNO, а сочетания типа *работать в Чикаго* - функцию LOC. Поэтому необходимо предусмотреть разную обработку сочетания ***работать клерком*** в случае, если у глагола РАБОТАТЬ нет зависимых типа ***в бухгалтерии*** или ***в Чикаго*** (подправило

1), и в случае, если такие зависимые имеются, например: ***работать в бухгалтерии клерком*** (подправило 2). Различие в обработке этих двух случаев состоит в том, что в первом из них функция JOB замещает глагол РАБОТАТЬ, а во втором эта функция подчиняется подлежащему. Сам глагол РАБОТАТЬ семантизируется тогда другим правилом.

5.4.4.7. Общее правило канонизации

Формальная запись правила

```
REG:CANON.7Ø        УСТРАНЕНИЕ ТАВТОЛОГИИ
TAKE:1
CHECK
1.1 =(X,FNC)&DOM-EQU(X,Z,ARG,VAR)&DOM-EQU(Z,Z1,JOT,FNC)&
     COLEX(X,Z1)&DOM-EQU(Z1,Z2,ARG,VAR)&COLEX(Z,Z2)
N:Ø1
CHECK
1.1 DEP(X,*,EQU)&DOM(Z1;W,EQU)&^DOM(Z2,*,*)
DO
1 IZSLOT:(X,*)-(W,*)
2 PERUZPREDNOM:W(X)
3 STERUZ:Z2
4 STERUZ:Z1
5 STERUZ:Z
6 STERUZ:X
N:Ø2
CHECK
1.1 DOM(X,W,EQU)&^DOM(Z1,*,EQU)
DO
1 IZSLOT:(X,*)-(Z1,*)
2 PERGRPREDNOM:Z1(X)
3 PERUZOT:(X,W,*)-(Z1,W,*)
4 PERGRSLEDNOM:W(Z2)
5 STERUZ:X
6 STERUZ:Z
```

Комментарий к правилу

Задача, которую решает это правило, подробно рассматривалась в предыдущем разделе (см. примеры (66)-(72)). Правило состоит из двух подправил, различающихся контекстом, в котором находится тавтологичное выражение.

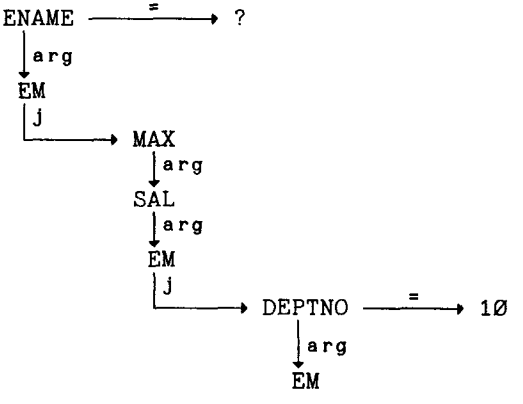
5.4.4.8. Трафаретное правило канонизации

Формальная запись правила

```
REG:CANON.11        РАСЩЕПЛЕНИЕ ПОДОПЕРАТОРНОЙ ФУНКЦИИ
TAKE:1              (ТРАФАРЕТНОЕ НА ОПЕРАТОРЫ MAX, MIN И AVG)
LOC:R
R:EQU/NON-EQU/MOR/LES/MOR-EQU/LES-EQU
N:Ø1
CHECK
1.1 ^DOM(X,*,R)&DOM-EQU(X,U,ARG,FNC)&DEP(X,Z,JOT)&
     DOM-EQU(U,Z1,ARG,VAR)&^DOM(U,*,R)
DO
1 DOBAUZ:W(FICT-LEX)
2 PERLEKAUZ:U(W)
3 SVUZOT:(Z,W,JOT)
4 PERUZPREDNOM:W(X)
5 DOBAUZ:W1(FICT-LEX)
6 PERLEKAUZ:Z1(W1)
7 SVUZOT:(W,W1,ARG)
8 PERUZSLEDNOM:W1(W)
9 PERUZOT:(Z,X,*)-(W,X,EQU)
1Ø IZGLOT:(X,CON)-(W,CON)
11 IZGLOT:(X,DIS)-(W,DIS)
12 IZREF:(Z,Z1)-(Z,W1)
```

Комментарий к правилу

Это правило применяется для обработки предложений типа ***Кто получает максимальную зарплату в отделе 10?*** (подробнее см. разд. 5.4.2.1, пример (29)). После семантизации это предложение получает следующую структуру:



Как мы видели в примере (29), такая структура передает смысл недостаточно эксплицитно. Для получения правильной структуры (29'), выражающей смысл '... служащие, зарплата которых равна максимальной зарплате служащих отдела 10', нужно расщепить функцию SAL, подчиненную оператору MAX.

5.4.4.9. Словарное правило канонизации

Формальная запись правила

```
REG:CANON.2Ø        ПЕРЕВЕШИВАНИЕ АРГУМЕНТА В НАЧАЛО
TAKE:1              СОЧИНИТЕЛЬНОЙ ЦЕПОЧКИ
LOC:R               (СЛОВАРНОЕ НА СЕМАНТИЧЕСКИЙ ЭЛЕМЕНТ '?')
R:CON/DIS
CHECK
1.1 DEP-EQUN(X,Z,EQU,FNC,OPR)&DOM-EQUN(Z,Z1,R,FNC,OPR)
1.2 DOM(Z1,W,ARG)/DOM(Z1,Z2,R)&DOM(Z2,W,ARG)
N:Ø1
CHECK
1.1 ~DOM(Z,*,ARG)
DO
1 PERUZOT:(*,W,ARG)-(Z,W,ARG)
2 PERGRSLEDNOM:W(X)
N:Ø2
CHECK
1.1 DOM(Z,U,ARG)&~DOM(U,*,*)&DOM(W,*,JOT)
DO
1 STERUZ·U
2 PERUZOT:(*,W,*)-(Z,W,*)
3 PERGRSLEDNOM:W(X)
```

Комментарий к правилу

Это правило предназначено для обработки конъюнктивно-дизъюнктивных цепочек из функций, которые имеют тождественные аргументы. Часто после семантизации аргумент имеется лишь у одной из них. Например, в предложениях типа ***Каковы фамилии и зарплаты клерков отдела сбыта?*** аргумент со значением 'клерки отдела сбыта' представлен только у функции SAL. Для последующего перевода на язык SQL удобно, чтобы общий аргумент находился при первой из функций, которая подчиняет вопросительный элемент '?'. Такое перевешивание и осуществляет данное правило.

5.5. Алгоритм семантического анализа

На вход этого алгоритма подается синтаксическая структура, представленная в виде дерева. На выходе получается семантическая структура, которая также является деревом. Процесс преобразования одной структуры в другую происходит в три основных этапа: нормализация, семантизация, канонизация. Каждый из этих этапов состоит из нескольких подэтапов, последовательность которых тоже заранее установлена. Преобразование дерева на каждом из подэтапов сводится к последовательному применению правил данного подэтапа.

Алгоритму должно быть заранее сообщено, какие правила подэтапа применяются вначале: общие или словарные и трафаретные. Обход узлов дерева также производится в определенном порядке: сверху (от вершины) вниз до максимальной глубины. При этом в каждой точке ветвления выбирается самый левый (из непройденных) путь.

Как легко видеть, алгоритм семантического анализа довольно прост: требуется шаг за шагом преобразовывать древесную структуру специально заготовленными последовательностями правил. Тем самым основная сложность этого этапа работы ЛП приходится на составление и упорядочение совокупности правил. Единственная нетривиальная ситуация касается правил, изменяющих структуру дерева.

В таких случаях обычно изменяется фрагмент, расположенный в дереве ниже узла, к которому применяется правило. Однако в некоторых правилах эти изменения касаются и фрагментов, расположенных выше этого узла. Такие изменения могут разрушить процедуру дальнейшего обхода дерева, организованную алгоритмом. Поэтому после работы с таким правилом следует вернуться к некоторому узлу дерева, выше которого изменения не происходили. В указанных случаях необходимая информация сообщается алгоритму в самом правиле посредством инструкции TAKE:Z, где Z - контекстная переменная, значение которой подбирается при проверке условий этого правила. Если обход дерева следует начать снова с вершины, используется инструкция TAKE:1. Алгоритм следит за тем, чтобы при таких возвратах одно и то же правило дважды не применялось.

Глава 6
СЛОВАРИ ЛИНГВИСТИЧЕСКОГО ПРОЦЕССОРА

В данной главе будут описаны два типа словарей, используемых в ЛП, - комбинаторный (КС) и семантический (СС). Морфологические словари русского и английского языков, используемые на этапах морфологического анализа и синтеза, были описаны в главе 3.

6.1. Комбинаторный словарь и его место в процессе переработки текста

В гл. 1 мы уже говорили о том, что первые два компонента ЛП - морфология и синтаксис - не ориентированы на конкретную прикладную задачу и конкретную проблемную область. Они задуманы как полифункциональные и могут быть использованы в широком спектре информационных систем, включая, в частности, системы МП. Чтобы эта возможность могла быть в принципе реализована, КС, в той мере, в какой он необходим для выполнения процедур анализа и синтеза текстов, должен быть существенно шире по объему в сравнении с тем набором слов, который нужен для системы общения с реляционными базами данных типа ORACLE. Этим объясняется то обстоятельство, что в КС включено большое количество слов, которые заведомо не могут встретиться в запросах к БД, однако очень часто встречаются в научно-технических текстах.

6.1.1. Типы информации о лексеме

Существенная особенность КС описываемого здесь ЛП состоит в разделении словарной информации на два типа - классификационную (предметную, термовую) и операционную, или информацию о правилах.

Необходимость такого разделения диктуется тем, что все правила системы разбиваются, как мы уже говорили, на три типа - общие, частные (трафаретные) и словарные. Общие правила, составляющие не более одной четверти всех правил, участвуют в обработке любой фразы. Частные и словарные правила участвуют в обработке лишь тех фраз, словесный матери-

ал которых может потребовать их применения. Они активируются словарными статьями соответствующих лексем, в которых они либо упоминаются по имени (трафаретные правила), либо приводятся целиком (словарные правила). Таким образом, в рассматриваемой версии ЛП реализуется идея оптимальной самонастройки системы на обработку текущего запроса. В результате время обработки запроса существенно сокращается, поскольку в его анализе принимают участие не все правила системы, а лишь те, в которых может возникнуть нужда.

Классификационная информация - это сведения о тех свойствах слова (часть речи, синтаксические и семантические признаки, стандартные способы оформления актантных зависимых в модели управления), на которые могут ссылаться правила системы. Операционная информация - это информация о трафаретных и словарных правилах.

6.1.2. Классификационная информация в словарной статье КС

6.1.2.1. Вход словарной статьи и описание многозначных слов

Словарная статья КС описывает ровно одну лексему (одно значение слова). Она открывается строкой заголовка, которая содержит номер словарной статьи и имя данной лексемы - слово в орфографической записи или безусловный оборот.

Напомним, что безусловным оборотом называется цепочка слов, выражающих единое понятие, имеющих неизменную грамматическую форму и следующих друг за другом в фиксированном порядке; ср. КАК БЫ ТО НИ БЫЛО, ВО ЧТО БЫ ТО НИ СТАЛО, ПО КРАЙНЕЙ МЕРЕ, ТЕМ НЕ МЕНЕЕ, КАК ЕСЛИ БЫ и т. п. Безусловным оборотам в КС присваивается метка определенной части речи.

Если слово имеет лексико-грамматические омонимы, то в имя каждой лексемы входит цифровой индекс, ср. ЧТО1 (союз) и ЧТО2 (союзное слово - *что, чего, чему* и т. п.). Если при этом надо различить еще и лексическую полисемию или омонимию, лексема получает двойной индекс вида i.j.

Отдельно необходимо сказать о том, как описываются в КС многозначные слова. Вообще говоря, лексическая многозначность, как и лексическая омонимия, создает серьезные технические затруднения для синтаксического анализа. Каждое значение слова и каждый омоним необходимо учитывать отдельно в качестве кандидата на положение хозяина или слуги в той или иной синтаксической конструкции (см. гл. 4), что приводит к

резкому росту числа гипотетических связей между словами фразы. Более точно: число гипотетических связей возрастает пропорционально квадрату числа омонимов во фразе. Поэтому даже в предложениях средней длины при наличии омонимов в достаточном числе позиций возникает такой комбинаторный взрыв, с которым машина технически может не справиться.

Из этого вытекает необходимость всеми возможными средствами, в том числе и искусственно, сокращать лексическую полисемию и омонимию. Естественный способ состоит в том, чтобы учитывать в КС не все возможные значения и лексические омонимы слова, а лишь те, которые релевантны для избранной предметной области. Искусственный же способ состоит в укрупнении значений, когда несколько разных значений объединяются в одно, даже если они имеют существенно разные синтаксические свойства.

Рассмотрим в качестве примера глагол БЫТЬ. Он представлен в русском КС в качестве единой лексемы. Между тем у него есть несколько разных значений: связочное (***Какова будет его зарплата?***), локативное (***Филиалы фирмы были в нескольких городах***), посессивное (***У каких служащих была максимальная зарплата?***), экзистенциальное (***Есть несколько аналитиков высокого класса, которые получают больше менеджеров***) и ряд других. Правильный семантический анализ таких словосочетаний легко обеспечить специальными правилами преобразования после того, как для предложения уже получена определенная синтаксическая структура. Самое же структуру следует строить в условиях минимальной омонимии, чтобы избежать неуправляемого комбинаторного взрыва. Именно такая стратегия была выбрана для лексикографической обработки многозначных слов в той версии КС, которая предназначена для ЛП и русско-английского перевода.

Классификационная информация отражается в четырех зонах словарной статьи (из которых обязательна только первая зона - зона части речи). Каждая зона вводится особой меткой: POR (часть речи), SYNT (синтаксические признаки), DES (семантические признаки), D1.j (совокупность строк, описывающих модель управления данной лексемы). В той версии КС, которая предназначена для русско-английского перевода, имеется зона тривиального перевода русской лексемы на английский язык, вводимая меткой TRANS

6.1.2.2. Зона части речи

В этой зоне после метки POR записывается символ одной из частей речи. Напомним (см. разд. 3.3), что мы выделяем следующие части речи: S (существительное), A (прилагательное), ADV (наречие), NUM (числительное), V (глагол), CONJ (союз), PR (предлог), PART (частица), COM (первый компонент сложных слов типа *франко-русский*). Дадим несколько содержательных пояснений.

В класс A, помимо канонических прилагательных, включены указательные местоименные прилагательные ЭТОТ, ТОТ; притяжательные местоименные прилагательные МОЙ, ТВОЙ, ВАШ, НАШ, вопросительные местоимения КАКОЙ, ЧЕЙ; неопределенные местоименные прилагательные типа ДРУГОЙ, НЕКОТОРЫЙ, ЛЮБОЙ, НИКАКОЙ, порядковые числительные типа ПЕРВЫЙ, ВТОРОЙ и т.п.; сложные адъективные местоимения типа КАКОЙ-ТО, КАКОЙ-НИБУДЬ, КАКОЙ-ЛИБО, КОЕ-КАКОЙ и слово НИЧЕЙ. Хотя это решение не соответствует традиционной точке зрения русской грамматики (см., напр., [Грамматика, 1960]), оно хорошо обосновано (особенно на русском материале) морфологически и синтаксически, имеет параллели в других грамматических традициях (например, во французской), а в последнее время было принято и в Академической грамматике русского языка (см. [Грамматика, 1970] и в особенности [Русская грамматика, 1980]).

Аналогичным образом расширен класс S, куда, кроме канонических существительных, включаются личные местоимения; вопросительные местоимения ЧТО2 и КТО; относительное местоимение КОТОРЫЙ; неопределенные местоимения НЕКТО и НЕЧТО; рефлексивное местоимение СЕБЯ; отрицательные местоимения НИКТО, НИЧТО; указательные местоимения ЭТО1 и ТО1; сложные субстантивные местоимения КОЕ-КТО, КОЕ-ЧТО, КТО-НИБУДЬ, КТО-ЛИБО, КТО-ТО, ЧТО-НИБУДЬ, ЧТО-ЛИБО, ЧТО-ТО.

В класс наречий, помимо собственно наречий, включены вводные слова и обороты типа ЕСТЕСТВЕННО, ВЕРОЯТНО, СЛОВОМ, ПО ВСЕЙ ВИДИМОСТИ, ПО ВСЕЙ ВЕРОЯТНОСТИ и т.п.; местоименные наречия типа ГДЕ, КОГДА2, ТАМ, ТУДА и т.п.; предикативы ("категория состояния") типа ЖАЛЬ, МОЖНО, НЕЛЬЗЯ, НУЖНО, НЕКОГДА, НЕГДЕ и т.п.; квантификаторы типа МАЛО, МНОГО, НЕМНОГО, НЕСКОЛЬКО, СКОЛЬКО, обычно относимые к числительным.

Все остальные части речи вполне соответствуют по объему традиционным понятиям.

6.1.2.3. Зона синтаксических признаков

Понятие синтаксического признака является весьма важным и в формальном, и в содержательном плане. К тому же как теоретическое понятие оно относительно ново (см., например, [Мельчук, 1974; Иомдин и др., 1975; Апресян, 1985; Mel'cuk, Pertsov, 1987]). Поэтому оно заслуживает подробного рассмотрения.

Под синтаксическими признаками понимаются сокращенные обозначения тех свойств слов, которые дают им возможность участвовать в одних синтаксических конструкциях и не допускают их участия в других. Хорошо известными грамматической традиции синтаксическими признаками являются род, *одушевленность/неодушевленность*, *исчисляемость/неисчисляемость* существительных, качественность, притяжательность и порядковость прилагательных, связочность и безличность глаголов, собирательность числительных, сочинительность и подчинительность союзов и т.п.

Ссылки на синтаксические признаки фигурируют во многих правилах синтаксического анализа и синтеза, а также в некоторых правилах других блоков.

Так, при анализе русских количественных конструкций необходимо проверять, обладает ли существительное - потенциальная вершина конструкции - признаком ИСЧИСЛ. Если да, то можно формировать гипотезу о возможности количественного отношения (количест) между этим существительным и содержащимся в предложении числительным. Если нет, то формировать такую гипотезу нельзя. Ср. словосочетание ***две большой важности проблемы***, для которого можно сформировать гипотезу

количест
┌──────────┐
две большой важности проблемы,

но не гипотезу

количест
┌──┐
две большой важности проблемы,

поскольку у существительного ПРОБЛЕМА признак ИСЧИСЛ есть (ср. ***две проблемы, пять проблем***), а у существительного ВАЖНОСТЬ - нет (ср. невозможность **две важности, *шесть важностей*).

Синтаксические признаки естественно сближаются с частями речи, потому что и те и другие задают грамматически значимые классы слов и возможные наборы их синтаксических функций. Однако между ними есть и существенные различия.

Части речи задают самые общие и нерасчлененные синтаксические функции весьма крупных по объему классов слов. Синтаксические признаки задают гораздо более узкие классы (иногда - всего по нескольку элементов в каждом) и гораздо тоньше детализируют синтаксические функции этих классов. Если некое слово является прилагательным, то из этого почти ничего не следует. Нельзя даже сказать, что такое слово может быть определением при существительном, потому что предикативные прилагательные, например, РАД или ДОЛЖЕН, как раз не могут быть определениями. Если же слову приписан какой-то синтаксический признак, то тем самым задается гораздо более содержательная информация о нем. Так, словам ЭТОТ, ТАКОЙ, ВСЕ и ряду других приписывается признак ПРИЧИСЛ. На этом основании они рассматриваются в качестве кандидатов на роль препозитивных определений к числительным; ср. конструкции типа *эти три терминала, такие два слова, все пять условий* и т.п. Очевидно, что далеко не все прилагательные могут употребляться в такой синтаксической позиции и функции; ср. неправильность конструкций **новые три терминала*, **длинные два слова*, **ошибочные пять условий*.

Каждое слово относится к одной и только одной части речи, между тем как синтаксических признаков у слова может быть более одного или не быть совсем. Так, у обозначений единиц измерения типа *градус, метр, ом,* час и т.д. есть уже упоминавшийся признак ИСЧИСЛ(яемость) и признак ИЗМЕР (единица измерения), нужный для анализа распределительных атрибутивных конструкций типа *по пять рублей метр.* У существительных - названий отрезков времени типа *час, день, неделя, месяц* и т.п. к этим двум признакам добавляются еще два: ДЛИТ(ельность) и ВРЕМ(я). Первый из них нужен для анализа длительных конструкций типа *работать три часа, находиться в плавании два месяца.* Второй нужен для анализа кратно-длительных и обстоятельственных конструкций типа *Он работал над каждой темой годами, Вечерами мы работали над книгой о лингвистическом процессоре.* Наконец, у многих при-

лагательных (в частности, большинства относительных), а также у ряда наречий, глаголов и предлогов нет никаких синтаксических признаков.

По указанной причине деление словаря по частям речи дает строгое разбиение, т.е. разбиение на непересекающиеся классы, а деление словаря по синтаксическим признакам - нет. Классы, задаваемые синтаксическими признаками, могут пересекаться и даже включаться друг в друга. Например, все русские слова, имеющие признак ВОПР (вопросительность, ср. ЧТО2, КАКОЙ, КАК, ПОЧЕМУ и пр.), имеют и признак МЕСТ (местоименность), т.е. целиком включаются в класс местоимений. Следовательно, в принципе русским лексемам, имеющим признак ВОПР, можно было бы и не приписывать признака МЕСТ: он автоматически выводится из признака ВОПР. Однако формально все синтаксические признаки трактуются как независимые, так, чтобы всякий раз можно было ссылаться на ту синтаксическую ипостась слова, которая в данном случае является содержательно релевантной. Более того, по той же причине мы считаем допустимым наличие в системе таких разных признаков, которые задают в точности один и тот же класс лексем, но характеризуют его с разных точек зрения. Таковы синтаксические признаки СВЯЗ (глаголы-связки типа БЫТЬ, СТАТЬ, ЯВЛЯТЬСЯ и т.п.) и СОГЛАКТ-2-1 (глаголы, требующие согласования второго актанта с первым по роду и числу, например, ***Перестройка стала необходимой, Этот признак оказался излишним, Эти меры были непопулярны***).

Еще одно различие между признаками и частями речи состоит в том, что синтаксические признаки могут приписываться словам, которые принадлежат различным частям речи. Так, признак МЕСТ приписывается существительным типа НИКТО, ЭТО и т.п.; прилагательным ЛЮБОЙ, ЭТОТ, КАЖДЫЙ и т.п.; наречиям ТАМ, ЗДЕСЬ, ГДЕ-ЛИБО и т.п. Признаки ЛОКАТ, КОНЕЧН, ИСХОДН приписываются наречиям и предлогам со значением, соответственно, локализации (ТАМ, В2, НА2, ПОД2, ЗА2 и т.п.), направления к конечной точке (ТУДА, СЮДА, В1, НА1, ПОД1, ЗА1 и т.п.) и направления из начальной точки (ОТТУДА, ОТСЮДА, ИЗ, С1, ОТ и т.п.). Признаки ПРЕДЕСЛИ, ПРЕДЧТО, ПРЕДВОПР, ПРЕДИНФ приписываются некоторым существительным, прилагательным и наречиям.

Наконец, в отличие от частей речи синтаксические призна-

ки во многих случаях семантически мотивированы; ср. уже упоминавшиеся признаки ДЛИТ, ИЗМЕР, ВРЕМ и ряд других. Однако в других случаях мотивация синтаксических признаков менее прозрачна. Поскольку стопроцентной семантической мотивацией синтаксические признаки не обладают, оказывается необходимым принципиально различать синтаксические и семантические признаки (см. описание зоны DES).

Обратим внимание еще и на связь между некоторыми синтаксическими признаками и морфологическими характеристиками. Мы имеем в виду такие грамматические категории, как род и одушевленность. В сущности, за каждой из них скрываются две разные категории: классифицирующая, в смысле [Зализняк, 1967], и словоизменительная. Для существительных они являются классифицирующими и их значения (МУЖСК, ЖЕНСК, СРЕДН, ОДУШ) трактуются нами как синтаксические признаки. Для прилагательных, причастий и глаголов они являются словоизменительными и трактуются как морфологические признаки, значениями которых являются характеристики (МУЖ, ЖЕН, СРЕД, ОД, НЕОД).

Ниже приводятся примеры русских синтаксических признаков, сгруппированных по частям речи. Другое возможное основание группировки - содержательный смысл синтаксических признаков. С этой точки зрения различаются признаки, описывающие типичную синтаксическую функцию слова (ср. АСПЕКТ-ВИН, ВОПР, ДЛИТ и т.п.), явления согласования (ЖЕНСК, КОЛЛЕК, ОДУШ, СОГЛАКТ-2-1, СОГЛАКТ-3-2 и т.п.), порядок слов (ПОСТОПР, ПРЕПОЗ!, ПРИЧИСЛ и т.п.), явления смещения (ДАТПОС, ПРЕДИНФ, СМЕЩ-1 и т.п.) и ряд других. Однако группировка признаков по частям речи не только практически полезнее, но и удобнее для обозрения.

Мы делаем упор на нетривиальные синтаксические признаки и их содержательный смысл, т.е. на лексико-семантические классы, для которых они характерны, и синтаксические конструкции, которые они обслуживают. Ни списки слов, которым приписывается тот или иной признак, ни списки обслуживаемых ими конструкций не являются (и не могут быть) исчерпывающими; они призваны иллюстрировать рассматриваемые синтаксические явления и показать, до какой степени глубины описываются синтаксические конструкции русского языка. Бытовой характер иллюстраций, возможно, вызовет недоумение; но мы

сознательно использовали общепонятные примеры, не привязанные ни к какой предметной области, чтобы сделать аппарат синтаксических признаков доступным для возможно более широкого круга специалистов, имеющих дело с системами лингвообработки.

1. Признаки глаголов

ВВОД-1 - глаголы, первая валентность которых заполняется предложением с союзом ЧТО, ЧТОБЫ и которые могут употребляться во вводной конструкции с союзом КАК: ВОДИТЬСЯ, ВЫЯСНЯТЬСЯ, КАЗАТЬСЯ, ОКАЗАТЬСЯ, ПОЛАГАТЬСЯ и т. п. ; ср. ***Некоторые сотрудники, как выяснилось в ходе расследования, не соблюдали элементарных правил техники безопасности.***

ВВОД-2 - глаголы, вторая валентность которых заполняется предложением с союзами ЧТО, ЧТОБЫ и которые могут употребляться в конструкции с союзом КАК: ВИДЕТЬ, ВЫЯСНЯТЬ, ГОВОРИТЬ, ДУМАТЬ, ОБЪЯВЛЯТЬ, ПОЛАГАТЬ, СЛЫШАТЬ, СЧИТАТЬ и т. п. ; ср. ***Некоторые сотрудники, как группа выяснила в ходе расследования, пренебрегали элементарными правилами техники безопасности.***

ГРАД - глаголы, обозначающие градуируемую ситуацию и способные присоединять наречия типа ОЧЕНЬ, ВЕСЬМА, НЕМНОГО, ЧУТЬ-ЧУТЬ и т. п. : ВЫРАСТИ, ЛЮБИТЬ, ОБИДЕТЬСЯ, УВЕЛИЧИТЬСЯ, УСТАВАТЬ, ХОТЕТЬ; ср. ***Дети очень выросли за лето, Он немного обиделся на меня.***

ДАТПОС - глаголы, способные присоединять в качестве неактантного комплетива в дательном падеже имя посессора своей нормальной валентности (ситуация смещения дополнения): ГЛАДИТЬ, ДУТЬ, ОТРЕЗАТЬ, ПРОЛИВАТЬ, СМОТРЕТЬ и т. п. ; ср. ***Дрессировщик посмотрел в глаза тигра → Дрессировщик посмотрел тигру в глаза.***

ДЕСТРУКТ - глаголы деструктивной деятельности, употребляющиеся с невалентными инструментальными обстоятельствами: ЕСТЬ, КОЛОТЬ, КРУШИТЬ, ЛОМАТЬ, РАЗБИВАТЬ, РАЗРУШАТЬ, РВАТЬ и т. п. ; ср. ***есть плов руками, крушить все топором, разбить окно локтем.***

ИЗМЕН - глаголы, обозначающие изменение состояния или положения в пространстве, способные сочетаться с обстоятельствами типа *на три дня:* ВСТАВАТЬ, ЕХАТЬ, ЗАМОЛЧАТЬ, ПРИХОДИТЬ, УЕЗЖАТЬ и т. п. ; ср. ***Машину остановили на три дня***

<совсем ненадолго> для профилактического ремонта, Он уехал на неделю в командировку. Ср. признак ИЗМЕН у существительных, имеющий совсем другой смысл.

КОНСТРУКТ - глаголы созидательной деятельности, употребляющиеся с вневалентными инструментальными обстоятельствами: ДЕЛАТЬ, ЕСТЬ, ПИТЬ, РАБОТАТЬ и т.п.; ср. *делать все своими руками, работать веслом.*

МГНОВ - глаголы мгновенного действия (так называемые моментальные), не способные сочетаться с обстоятельствами длительности типа *долго, недолго, три дня, неделю, (целый) месяц:* ВЕЛЕТЬ, ДОСТИГАТЬ, КАСАТЬСЯ, НАХОДИТЬ, ОСТАВЛЯТЬ, ПРИХОДИТЬ и т.п.; ср. **Судно целый день приходило в порт,* **Альпинисты два дня достигали вершины.*

МОД - модальные глаголы: МОЧЬ, УМЕТЬ, ХОТЕТЬ, способные подчинять цепочку последовательно подчиненных инфинитивов, ср. *Завод может перестать работать в любой момент из-за нехватки сырья.*

НЕСОВ! - глаголы, не имеющие формы совершенного вида: БЫТЬ, ВЕЗТИ, ЗНАТЬ, ИДТИ, НАСТУПАТЬ (об армии), СПАТЬ и др.

НЕ-ЧТОБЫ - глаголы, которые прототипически управляют предложением с союзом ЧТО, но в контексте отрицания способны менять ЧТО на ЧТОБЫ: ВЕРИТЬ, ВИДЕТЬ, ДУМАТЬ, СЛЫШАТЬ и т.п.; ср. *Я слышал, что он собирается на пенсию → Я не слышал, чтобы он собирался на пенсию.*

ОКОПР - переходные глаголы, употребляющиеся в объектно-копредикативной конструкции: ВИДЕТЬ, ВСТРЕЧАТЬ, ДОСТАВЛЯТЬ, ЗНАТЬ, ПОМНИТЬ, ПРИВЕЗТИ и т.п.; ср. *Мы знали его* [объект] *молодым* [копредикатив, согласованный с объектом в роде и числе], *Машины были доставлены совершенно исправными.*

РДОП - переходные глаголы, способные менять винительный падеж дополнения на родительный в контексте отрицания: ДЕЛАТЬ, ЕСТЬ, ПИСАТЬ, ПИТЬ, ЧИТАТЬ и т.п. (подавляющее большинство переходных глаголов); ср. *Все читали эту книгу → Никто не читал этой книги.*

РПОДЛОТР - глаголы, которые допускают мену именительного падежа подлежащего на родительный в контексте отрицания: БЫВАТЬ, БЫТЬ, ВОДИТЬСЯ, ВЫДЕЛЯТЬСЯ, ИМЕТЬСЯ, ПОЛАГАТЬСЯ, ПОСТУПАТЬ, СУЩЕСТВОВАТЬ, ТРЕБОВАТЬСЯ и т.п., ср. *Такие сведения* [ИМ] *в институт еще не поступали - Таких сведений* [РОД] *в институт еще не поступало.*

СВЯЗ - связочные глаголы БЫТЬ, БЫВАТЬ, ДЕЛАТЬСЯ, КАЗАТЬСЯ, ОКАЗЫВАТЬСЯ, ОСТАВАТЬСЯ, ПРЕДСТАВЛЯТЬСЯ, СТАНОВИТЬСЯ, СЧИТАТЬСЯ, ЯВЛЯТЬСЯ.

СКОПР - глаголы, употребляющиеся в субъектно-копредикативной конструкции: ВИСЕТЬ, ЖИТЬ, ЗНАТЬ, ПРИЕЗЖАТЬ, ПРИСНИТЬСЯ, РОДИТЬСЯ; ср. *Отец* [субъект] ***знал ее молодым*** копредикативный член, согласованный с субъектом в роде и числе].

СОВ! - глаголы, не имеющие формы несовершенного вида: ПОЙТИ и другие производные на ПО- от глаголов перемещения, ВЕРНУТЬСЯ, РАЗГЛЯДЕТЬ (РАЗГЛЯДЫВАТЬ - другая лексема) и т. п.

*СОВ-1 - глаголы, которые не способны управлять на первом месте инфинитивом совершенного вида (при допустимости несовершенного): НАДОЕСТЬ, НАСКУЧИТЬ, ОПРОТИВЕТЬ, ПОНРАВИТЬСЯ, ХВАТИТ и т. п.; ср. ***Ему надоело лгать.***

СОГЛАКТ-2-1 - глаголы, требующие согласования второго актанта с первым по числу и роду: БЫТЬ, ДЕЛАТЬСЯ, КАЗАТЬСЯ, СТАНОВИТЬСЯ, ЯВЛЯТЬСЯ и другие связки, такие, как ОКАЗЫВАТЬСЯ, ПРЕДСТАВЛЯТЬСЯ, ПРИКИДЫВАТЬСЯ, ПРИТВОРЯТЬСЯ, СЧИТАТЬСЯ и т. п. Ср. *План* [ЕД, МУЖСК, 1-й актант] ***кажется*** <***считается, представляется***> ***нереальным*** [ЕД, МУЖ, 2-й актант], *Программа* [ЕД, ЖЕНСК, 1-й актант] ***кажется*** <***считается, представляется***> ***нереальной*** [ЕД, ЖЕН, 2-й актант], *Планы* [МН, 1-й актант] ***кажутся*** <***считаются, представляются***> ***нереальными*** [МН, 2-й актант]; здесь МУЖСК и ЖЕНСК - синтаксические признаки, а МУЖ и ЖЕН - морфологические характеристики.

СТР - глаголы, допускающие образование страдательного причастия совершенного вида: ДЕЛАТЬ, ПИСАТЬ, ПОЛУЧАТЬ, ЧИТАТЬ и т. п., но не СЛУШАТЬ (***меня внимательно***), ЧУВСТВОВАТЬ (***страх***); ср. ***сделанный, написанный, полученный, прочитанный*** и т. п., но не **послушанный, *почувствованный.*

СТР-СЯ - глаголы, допускающие образование синтетической формы страдательного залога: ДЕЛАТЬ, ПИСАТЬ, СТРОИТЬ, но, например, не ВИДЕТЬ, ЛЮБИТЬ, ПОЛУЧАТЬ, СЛЫШАТЬ, СЧИТАТЬ; ***Все делается ими без спешки, Масса бумаг пишется референтами, Мост строился солдатами, *Нами любятся прогулки, *Кем получаются письма?, *Иваном слышатся какие-то странные звуки, *Иван считается мной талантливым.***

СТР-ЕМ - глаголы, допускающие образование страдательного причастия несовершенного вида: ОБУЧАТЬ, ПОЛУЧАТЬ, ПРЕПОДАВАТЬ, РАЗРУШАТЬ и т.п., но, например, не СТРОИТЬ, СЧИТАТЬ; ср. *обучаемый, получаемый, преподаваемый, разрушаемый* при невозможности **строимый, *считаемый (Иван, считаемый мной талантливым)*.

Признаки серии СТР неожиданно оказались продуктивными при синтезе правильной формы страдательного залога для данного глагола, потому что сами глаголы весьма причудливо распределены относительно трех различных форм пассива. В частности, далеко не все формально переходные глаголы обладают всеми формами пассива; так, ПРИНИМАТЬ в значении 'получать' имеет все три формы (*принятый, принимается, принимаемый*), а синонимичный ему глагол ПОЛУЧАТЬ - только две (*полученный, получаемый*, но не **получается* - см. выше).

ФАЗ - фазовые глаголы: КОНЧАТЬ, НАЧИНАТЬ, ПРЕКРАЩАТЬ, ПРОДОЛЖАТЬ, СТАТЬ и т.п.

2. Признаки существительных

АГЕНС - существительные, способные быть агентивным дополнением при пассивной форме глагола, например: ЧЕЛОВЕК, АЛГОРИТМ, СОЛНЦЕ и т.п.; ср. *Запрос обрабатывается компьютером <оператором>*.

АДЪЕКТ - существительные адъективного склонения типа БУДУЩЕЕ, КОМАНДИРОВОЧНЫЕ, МОСТОВАЯ, НАСЕКОМОЕ, РАБОЧИЙ, СТОЛОВАЯ, ЧАСОВОЙ, ЗЛОТЫЙ (для обеспечения правильной формы числа в малых количественных группах, в которых "нормальные" существительные имеют форму ЕД, а существительные адъективного склонения - форму МН; ср. *два рубля* [ЕД], но *два злотых* [МН]).

АСПЕКТВИН - параметрические существительные типа ВЫСОТА, ГЛУБИНА, ДЛИНА, ТОЛЩИНА, ШИРИНА и т.п., способные выступать в качестве атрибута в винительном падеже в конструкциях типа *около ста метров в высоту <в длину, в ширину>*.

АТР - существительные со значением атрибутов человека - частей тела, предметов одежды, болезней и т.п.: ГОЛОВА, ГРИПП, НОС, ОЧКИ, ТРУБКА, ШЛЯПА и т.п. Они выступают в разного рода неактантно-комплетивных и копредикативных конструкциях типа *заглянуть тигру в глаза, увидеть Ивана в шляпе, Петр пришел на работу с ангиной*.

ВРЕМ - существительные со значением точки во времени или временного промежутка: ВЕК, ДЕНЬ, МИНУТА, ПОРА, ЧАС, ЭПОХА и т. п.; обслуживает все длительные конструкции, а также некоторые атрибутивные и обстоятельственные; ср. *работать неделю, просиживать часами, Россия в двадцатом веке, пять дней спустя; с того знаменательного дня, как мы познакомились.*

ГЕОГР - существительные со значением географического объекта, не являющиеся именами собственными: ГОРА, ГОРОД, ДЕРЕВНЯ, РЕКА, СТРАНА и т. п.; обслуживает аппозитивные конструкции типа *город Великий Устюг.*

ГИПЕР - существительные со значением родового понятия: АЛЛЮР (ср. *рысь, галоп, карьер*), ДЕНЬ (ср. *понедельник, вторник,..*), МЕСЯЦ (ср. *январь, февраль,..*), СТИЛЬ (ср. *баттерфляй, кроль,..*), МЕТОД, СПОСОБ и т. п.; соответствующим видовым названиям не приписывается; обслуживает атрибутивные и обстоятельственные конструкции типа *обработка методом квадратов, встреча в следующем месяце, обрабатывать методом квадратов, встречаться в следующем месяце* (но не **обработка методом, *встреча в месяце, *обрабатывать методом, *встречаться в месяце*; ср., с другой стороны, *обработка давлением, встречаться в феврале* и т. п.).

ДЕН - существительные - названия денежных единиц, способные вводить распределительную группу без предлога: ДОЛЛАР, КОПЕЙКА, МАРКА, РУБЛЬ, ФРАНК, ФУНТ и т. п.; ср. *компьютеры по 40 000 рублей штука.*

ДРОБ - существительные, обозначающие дробные доли и не способные употребляться в аппроксимативно-количественных конструкциях: ВТОРАЯ, ТРЕТЬЯ, ЧЕТВЕРТАЯ, ПЯТАЯ и т. п.; ср. *три четвертых, четыре пятых,* но не **четвертых три, *пятых четыре* в отличие от "нормальных" существительных типа КНИГА, ПРОЕКТ, СТРАНИЦА и т. п.: *книг пять, проектов десять, страниц сорок.*

ЖЕНСК - существительные женского рода.

ИЗМЕР - названия единиц измерения типа ГРАММ, ГРАДУС, ДЕСЯТОК, МЕТР, РУБЛЬ, ЧАС и т. п.

ИЗМЕН - существительные, способные присоединять атрибут или обстоятельство типа *в n раз*: ИЗМЕНЕНИЕ, РОСТ, СОКРАЩЕНИЕ, УВЕЛИЧЕНИЕ, УМЕНЬШЕНИЕ и т. п.; ср. *рост <сокращение, увеличение> производства в два раза.*

ИСЧИСЛ - исчисляемые существительные, способные употребляться в количественных и аппроксимативных группах; ср *пять рублей, километров пять-шесть.*

КЛАС - существительные, обозначающие классы, роды, виды и параметрические свойства: ВИД, КЛАСС, КРУГ, РОД, ТИП,... ГАБАРИТЫ, КАЛИБР, РАЗМЕР,...; КРАСОТА, МУЖЕСТВО, УМ, ХАРАКТЕР и т.п. Для них характерна способность употребляться в атрибутивных конструкциях типа *объекты этого класса, орудия большого калибра, птицы отряда воробьиных.*

КОЛЛЕКТ - существительные типа БОЛЬШИНСТВО, МЕНЬШИНСТВО, РЯД, ЧАСТЬ и т.п., по-особому согласующиеся со сказуемым, ср. *Большинство компьютеров из этой партии оказались непригодными к эксплуатации - Большинство сотрудников не пришло на собрание.*

КОНТЕЙН - существительные со значением вместилища (контейнера), способные употребляться в субъектно-обстоятельственных и объектно-обстоятельственных конструкциях: БОЧКА, ВЕДРО, ЛОЖКА, СТАКАН, ЯЩИК и т.п.; ср. *Птицы селились целыми стаями* [субъектное обстоятельство], *Они ели апельсины ящиками* [объектное обстоятельство]. В связи с этим признаком заметим, что значения субъекта и объекта выражаются не только подлежащими и дополнениями и не только копредикативными членами, но и многими видами обстоятельств; поэтому распознавание соответствующих синтаксических конструкций оказывается важным для дальнейшего семантического анализа высказываний.

ЛИЧН - личные местоименные существительные типа Я, ТЫ, ОН, МЫ и т.п., имеющие более сложные правила согласования, чем обычные существительные, и поэтому рассматриваемые в синтагмах особо.

ЛОК - существительные, имеющие местный падеж: БОР, ВЕТЕР, МОЛ, МОСТ, СНЕГ и т.п.; ср. *На мосту было ветрено* [ср. *О Бруклинском мосте написаны поэмы*], *На бору со звонами плачут глухари* (С.Есенин).

МЕС - названия месяцев, способные употребляться в обстоятельствах даты и в аппозитивных конструкциях особого типа: ЯНВАРЬ, ФЕВРАЛЬ и т.п., но не само слово МЕСЯЦ (ср. признак ГИПЕР); *Мы приступили к работе 30 мая <в мае месяце>.*

МН! - существительные pluralia tantum (имеющие только форму множественного числа).

МУЖСК - существительные мужского рода.

ОБРАЩ - существительные, выступающие в роли стандартных обращений и способные употребляться в ряде аппозитивных конструкций: ВЕЛИЧЕСТВО, ГОСПОДИН, ГРАЖДАНИН, ЛЕЙТЕНАНТ, ТОВАРИЩ и т.п.; ср. *гражданин Иванов.*

ОДУШ - одушевленные существительные типа ВАЛЕТ, ВОЛК, ДИРЕКТОР, МЕРТВЕЦ, ПЕТЯ, РОБОТ, ФЕРЗЬ, ЧЕЛОВЕК, а также КТО, СЕБЯ и т. п., имеющие ряд синтаксических особенностей в конструкциях типа *читать умных авторов* [ОДУШ, винительный падеж совпадает с родительным] - *читать умные книги* [не ОДУШ, винительный падеж совпадает с именительным], *заглянуть тигру* [ОДУШ] *в глаза* (возможная конструкция) - **заглянуть дому* [не ОДУШ] *в окна* (невозможная конструкция; надо *заглянуть в окна дома*), *трое операторов* [ОДУШ, МУЖСК, возможная конструкция) - **трое компьютеров* [не ОДУШ, не МН!, невозможная конструкция; надо: *три компьютера*].

ОЦЕН - оценочные существительные типа ДУРАК, ЗАНУДА, МОЛОДЕЦ, ТИРАН, УМНИЦА и т.п., способные употребляться в аппозитивной конструкции типа *начальник тиран.*

ПАРАМ - параметрические существительные типа ВЕС, ВЫСОТА, ДЛИНА, МАССА, РАЗМЕР, ТОЛЩИНА, УРОВЕНЬ, ФОРМА, ЦВЕТ, ЧАСТОТА, ШИРИНА и т.п., формирующие атрибутивную конструкцию типа *Подъездные пути большой длины <протяженности>.*

ПАРТИТ - уменьшительные существительные типа БЕНЗИНЧИК, КЕФИРЧИК, ЛЕДОК, МЕДОК, САХАРОК, ЧАЕК, которые употребляются в партитивном падеже даже в тех случаях, когда соответствующие неуменьшительные существительные нормально употребляются в родительном, ср. *пачка <чашка> чая*, но только *пачка <чашка> чайку* при невозможности **пачка чайка.*

ПОСТОПР - местоименные существительные типа ВСЕ, КТО-ТО, КТО-ЛИБО, ЧТО, ЧТО-НИБУДЬ и т.п., требующие постпозиции определения; ср. *кто-то незнакомый, что-то интересное.*

ПРОФ - существительные мужского рода, являющиеся названиями профессий и допускающие при себе сказуемое или определение в женском роде: ВРАЧ, ДИРЕКТОР, ПИЛОТ, ПРОДАВЕЦ и т.п.; ср. *В комнату вошла администратор гостиницы.*

РАСПР - существительные, способные на втором месте подчинять "денежную" распределительную группу в именительном падеже: СТОИМОСТЬ, ТАКСА, ТАРИФ, ЦЕНА и т.п.; ср. *по цене два рубля (за) килограмм.*

СЕЗ - названия времен года, которые в форме творительного падежа способны выступать в роли временного атрибута или обстоятельства: ВЕСНА, ЗИМА, ЛЕТО, ОСЕНЬ; ср. *Москва весной, Он приехал поздней осенью.*

СИМВ - существительные, являющиеся названиями знаков и способные употребляться в аппозитивных конструкциях: БУКВА, ЗНАК, ИМЯ, СЛОВО, ФАМИЛИЯ, ФРАЗА, ЦИФРА и т. п.; ср. *латинская буква q.*

СМЕЩ-i [СМЕЩ-1, СМЕЩ-2 и т. д.] - существительные, способные передавать свой i-й актант глаголу функционального типа, у которого таким образом появляется смещенное дополнение, не отвечающее никакой семантической валентности этого глагола: ВОЗРАЖЕНИЕ, ПОМОЩЬ, ПРЕДЛОЖЕНИЕ, СВЯЗЬ и т. п.; ср. *Возражение* [СМЕЩ-1] *поступило со стороны министра* [министр - первый актант ситуации возражения]; *Ивану оказали помощь* [СМЕЩ-2; Иван - второй актант ситуации помощи, то есть тот, кому помогли; первым является тот, кто помог]; *За поимку преступника он получил крупное вознаграждение* [СМЕЩ-3; поимка - третий актант ситуации вознаграждения; первые два - тот, кто вознаграждает, и тот, кого вознаграждают]; *дать на два года гарантию* [СМЕЩ-4; срок - четвертый актант ситуации гарантии; первые три - тот, кто гарантирует, тот, кому гарантируют, и то, что гарантируют].

СОВОК - существительные со значением совокупности (в том числе все КОЛИЧ и все КОНТЕЙН), способные формировать особого рода предикативные, копредикативные и некоторые другие синтаксические конструкции: БОЧКА, ГРУППА, ЗАВОД, ПАРТИЯ, СТАДО, ТЫСЯЧА, ЯЩИК и т. п.; ср. *Огурцов привезли целую бочку <тонну>, Столовые получали апельсины маленькими партиями, Коров было огромное стадо.*

СРЕДН - существительные среднего рода.

ТВОРАТР - непараметрические существительные, которые в форме творительного падежа могут употребляться в значении способа осуществления действия и выполнять функцию атрибута или обстоятельства: БРАСС, МЕТОД, ОБРАБОТКА, ПЛАВАНИЕ, СВАРКА, СПОСОБ и т. п.; ср. *сеять что-л. <посев чего-л.> гнездовым методом <способом>, сшивание металлических листов сваркой.*

ТВОРОБСТ - существительные со значением времени, способ-

ше в творительном падеже выступать в функции обстоятельства: ВЕСНА, ВРЕМЯ, ЗИМА, ЛЕТО, ОСЕНЬ, ПОРА и т. п.; ср. *приехать зимой, встречаться летней порой.*

ТРАНСП - существительные, являющиеся родовым или нейтральным именем транспортного средства и способные обозначать соответствующий тип (а не инструмент) перемещения: АВТОБУС (но не "ИКАРУС"), ГРУЗОВИК (но не "МАЗ"), МАШИНА (но не ЛИМУЗИН), ПАРОХОД (но не ЛАЙНЕР), ПОЕЗД (но не "КРАСНАЯ СТРЕЛА"), САМОЛЕТ (но не ЛАЙНЕР), ТРАМВАЙ, ТРОЛЛЕЙБУС и т. п.; ср. *доехать поездом, прилететь самолетом, ездить на работу трамваем <автобусом>.*

ФПРЕДВОПР - существительные, которые при соединении с полувспомогательным глаголом типа OPER или FUNC индуцируют у него способность иметь в качестве подлежащего вопросительное придаточное предложение: БЕСПОКОЙСТВО, БОЯЗНЬ, КОЛЕБАНИЯ, НЕУВЕРЕННОСТЬ, СОМНЕНИЕ и т. п.; ср. *Вызывает сомнение, способен ли он на бескорыстный поступок.*

ФПРЕДИНФ - существительные, которые при соединении с полувспомогательным глаголом типа OPER или FUNC индуцируют у него способность иметь в качестве подлежащего инфинитив: ОБИХОД, ОБЫЧАЙ, ПРАВИЛО, ПРИВЫЧКА и т. п.: *Он взял за правило делать по утрам гимнастику, У него вошло в привычку делать по утрам гимнастику.*

ФПРЕДКОГДА, ФПРЕДЧТО, ФПРЕДЧТОБЫ - то же самое, но с подлежащим - придаточным предложением, вводимым союзами КОГДА, ЧТО или ЧТОБЫ соответственно; ср. *Мы примем к сведению, что заведующие будут отсутствовать.*

ХАРАКТПР - существительные, способные выступать в роли атрибута в предложном падеже с обязательным зависимым при атрибуте: ВЕС, ВКУС, ДУХ, ИЗОБРАЖЕНИЕ, ИСПОЛНЕНИЕ, ПЕРЕВОД, ПЕРЕДАЧА, РОЛЬ, СТЕПЕНЬ, ФОРМА, ЧИН и т. п.; ср. *боксеры в весе до 70 кг., пьеса во вкусе Ионеско, Данте в переводе Лозинского, военный в чине полковника.*

ХАРАКТРОД - существительные, способные выступать в роли атрибута в родительном падеже с обязательным зависимым при атрибуте: ВИД, ГИЛЬДИЯ, ГРУППА, ДЕЙСТВИЕ, ДОБРОТА, ЖАНР, ЗОЛОТО, КАЧЕСТВО, МАНЕРА, МРАМОР, ПРОБА, СТИЛЬ и т. п.; ср. *предложения такого вида, купец первой гильдии, человек необычайной доброты, браслет червонного золота, золото высшей пробы.*

ХАРАКТТВОР - существительные, способные выступать в роли атрибута в творительном падеже с обязательным зависимым при атрибуте: МЕТОД, ОБРАЗ, ПРИЕМ, ПУТЬ, СПОСОБ, СТИЛЬ и т.п., а также названия параметров типа ВЫСОТА, ДЛИНА, МОЩНОСТЬ, РАЗМЕР и т.п.; ср. *мотор мощностью в 150 л.с., решать задачу двумя способами, сеять морковь методом квадратов.*

3. Признаки прилагательных

КВАНТЪЕД-МУЖ - прилагательные, способные в положительной степени в мужском роде выступать в роли субстантива: ВСЯКИЙ, ЕДИНСТВЕННЫЙ, КАЖДЫЙ, ПЕРВЫЙ, ПОСЛЕДНИЙ и т.п.; ср. *Всякий, кто жил в Италии, знает...; Единственный, кого он помнил, был Иван.*

КВАНТЪЕД-ПРЕВ - прилагательные, способные в синтетической превосходной степени среднего рода выступать в роли субстантива: ВАЖНЕЙШЕЕ (от ВАЖНЫЙ), ЛУЧШЕЕ (от ХОРОШИЙ), УМНЕЙШЕЕ (от УМНЫЙ), ХУДШЕЕ (от ПЛОХОЙ) и т.п.; ср. *Лучшее, на что он способен <умнейшее, что он мог сделать>, это не прийти на собрание.*

КВАНТЪЕД-СРЕД - прилагательные, способные в положительной степени среднего рода выступать в роли субстантива: ГЛАВНОЕ, ЕДИНСТВЕННОЕ, ОСНОВНОЕ, ПЕРВОЕ, ПОСЛЕДНЕЕ и т.п.; ср. *Единственное, что его волнует, это его собственное благополучие.*

ОЦЕН-РЕЛЯТ - прилагательные, которые могут употребляться в форме среднего рода в релятивных конструкциях в роли субстантива: БЛАГОРОДНЫЙ, ГЛУПЫЙ, ДОРОГОЙ, ИНТЕРЕСНЫЙ, НЕОБХОДИМЫЙ, ПЛОХОЙ и т.п.; ср: *Самое интересное <доступное, хорошее>, что они могли бы нам предложить.* В этих случаях от них должно зависеть кванторное слово типа ВСЕ (*все интересное, что она могла сказать*), САМЫЙ, НАИБОЛЕЕ и т.п.

ПОРЯДК - порядковые прилагательные, в том числе обозначенные цифрами, способные употребляться в большом круге аппроксимативно-количественных, атрибутивных, обстоятельственных и длительных конструкций: ПЕРВЫЙ, ВТОРОЙ, ТРЕТИЙ, ..., ДЕСЯТЫЙ, ..., ДВАДЦАТЬ ПЯТЫЙ и т.п.; ср. *День пятый-шестой считается самым трудным для акклиматизации, Полиция догнала нарушителя километре на двадцатом, встреча в верхах 7 декабря, Каждый четвертый житель городка получит в текущем году новую квартиру.*

ПРИТЯЖ - притяжательные прилагательные, занимающие в цепочках неоднородных адъективных определений, состоящих из относительных прилагательных, крайнее левое место (перед ними могут стоять только кванторные или указательные определения или определения в превосходной степени); БОРИН, МОЙ, НАШ, ОТЦОВ и т. п.; ср. ***бабушкина швейная машина,*** но ***каждое мое слово, эта ваша привычка, лучшее Катино платье.***

ПРИЧИСЛ - прилагательные, нормально употребляющиеся в роли препозитивных определений к числительным: ВСЕ, ТАКОЙ, ЭТОТ и т. п.; ср. ***все двенадцать кресел*** (не ****двенадцать все кресел***).

ПРОСТР - прилагательные со значением локализации в пространстве, способные выступать в качестве рестриктивного определения в конструкции со смещенным дополнением при невозможности квалификативного определения: ВЕРХНИЙ, ЗАДНИЙ, ЛЕВЫЙ, НИЖНИЙ, ПЕРЕДНИЙ, ПРАВЫЙ и т. п.; ср. ***вылечить псу заднюю лапу*** при невозможности ****вылечить псу мохнатую лапу.***

СИНТЕТ - прилагательные, имеющие синтетическую сравнительную степень: БОЛЬШОЙ, ВЫСОКИЙ, УМНЫЙ, ХОРОШИЙ и т. п.

4. Признаки наречий

ВВОД - вводные наречия и наречные обороты типа ВЕРОЯТНО, ВО-ВТОРЫХ, ВОЗМОЖНО, ДОЛЖНО БЫТЬ, КОНЕЧНО, КРОМЕ ТОГО, МОЖЕТ БЫТЬ, ПРИЗНАТЬСЯ, РАЗУМЕЕТСЯ и т. п., способные формировать вводные конструкции; ср. ***Он, вероятно, уже приехал.***

В-РАЗ - наречия, способные заполнять ту валентность глаголов с признаком ИЗМЕН (РАСТИ, УВЕЛИЧИВАТЬСЯ, УМЕНЬШАТЬ), которая характеризует величину изменения: ВДВОЕ, ВТРОЕ, ВЧЕТВЕРО и т. п.; ср. ***Объем производства вооружений сократился вдвое.***

ДЕБИТ - наречия, не сочетающиеся с глаголами совершенного вида прошедшего времени: ВО ЧТО БЫ ТО НИ СТАЛО, НЕПРЕМЕННО, ОБЯЗАТЕЛЬНО и т. п.; ср. неправильность ****Он обязательно написал вам письмо.***

ИНТЕНС - наречия степени, способные употребляться в роли ограничительных зависимых в сочетании с качественными прилагательными и в ряде других синтаксических конструкций: АБСОЛЮТНО, БОЛЕЕ, ВПОЛНЕ, ДОВОЛЬНО, ДОСТАТОЧНО, МЕНЕЕ, НАИМЕНЕЕ, ОЧЕНЬ, ПОЛНОСТЬЮ, СЛИШКОМ, СОВЕРШЕННО, СОВСЕМ и т. п.; ср. ***довольно интересный, чересчур тяжелый.***

МАНЕР - наречия образа действия, способные формировать стандартные обстоятельственные конструкции: БЫСТРО, ЛЕГКО, ПОЛНОСТЬЮ, ПОСТЕПЕННО, РЕЗКО, ТЯЖЕЛО и т. п.; ср. ***Добыча нефти в стране существенно увеличилась.***

МНОЖ - наречия, требующие наличия в предложении идеи множественности либо в виде формы МН, либо в виде кванторных слов типа ВСЕ, либо в виде сочинительной или комитативной (с предлогом СЗ) группы: ВМЕСТЕ, ВРОЗЬ, ЗАОДНО, ОДИНАКОВО, ПО-РАЗНОМУ, СООБЩА и т. п.; ср. ***Они <все> пришли к финишу одновременно, Француз и голландец <француз с голландцем> пришли к финишу одновременно.***

ОБ-МНОЖ - наречия, требующие объекта либо во множественном числе, либо в виде кванторного местоимения типа ВСЕ, либо в виде сочинительной группы в синтаксической позиции объекта: НАРАСХВАТ, ПОИМЕННО и т. п.; ср. ***перечислить кандидатов <всех> поименно, назвать поименно мастера и гроссмейстера, участвовавших в том матче.***

ПРЕД! - наречия, способные выступать только в предикативной функции при некоторых связках: В СИЛАХ, ЖАЛЬ, МОЖНО, НАДО, НЕЛЬЗЯ, НУЖНО и т. п.; ср. ***Помочь вам мы не в силах, Расставаться с ним было жаль.***

ПРЕДПАРТ - наречия, индуцирующие у связочного глагола, от которого они зависят, подлежащее в партитиве или в родительном падеже: БОЛЬШЕ, ДОСТАТОЧНО, МАЛО, МЕНЬШЕ, МНОГО, НЕДОСТАТОЧНО, НЕМАЛО, НЕМНОГО, СКОЛЬКО, СТОЛЬКО и т. п.; ср. ***Хлеба было мало <недостаточно>, Людей на площади оказалось больше <меньше>, чем вчера.***

ПРИГЛАГ! - наречия, которые могут зависеть только от глаголов: БЫСТРО, ДВАЖДЫ, ДЕЙСТВЕННО, ОДНАЖДЫ, ОКОНЧАТЕЛЬНО, ПОЭТОМУ, ТРИЖДЫ, УСПЕШНО и т. п.; ср. ***Мы быстро наверстаем упущенное, Лектор дважды возвращался к этому вопросу.*** Наречия ДВАЖДЫ, ТРИЖДЫ и т. п. могут употребляться также в конструкции типа ***дважды <трижды> Герой Советского Союза,*** в которой они зависят от существительных. Мы, однако, не учитываем эту конструкцию ввиду малой вероятности ее появления в рассматриваемых нами текстах.

ПРИПОРЯДК - пространственные наречия, способные атрибутивно присоединяться к порядковым прилагательным: НАЛЕВО, НАПРАВО, ПРЯМО, СВЕРХУ, СЛЕВА, СНИЗУ и т. п.; ср. ***вторая улица направо, первый дом слева.***

ПРИПРИЛ! - наречия, которые могут зависеть только от прилагательных и наречий: ВЕСЬМА, ГОРАЗДО, ДОВОЛЬНО, ИСКЛЮЧИТЕЛЬНО и т. п.; ср. ***весьма интересный, гораздо больше.***

ПРИСРАВН - наречия, которые могут зависеть от сравнительной степени прилагательных и наречий: ВДВОЕ, ВПОЛОВИНУ, ВТРОЕ, ЗНАЧИТЕЛЬНО, НАМНОГО, СУЩЕСТВЕННО, ЧУТЬ-ЧУТЬ и т. п.; ***Конкурирующая фирма продвинулась значительно дальше в своей разработке, Ваша сестра намного красивее.***

ПРИСРАВН! - наречия, которые могут употребляться только при наречиях и прилагательных в сравнительной степени: ГОРАЗДО, НАМНОГО и т. п.; ср. ***Она гораздо <намного> красивее.***

ПРИСУЩ - наречия, способные присоединяться к существительным в качестве атрибута: ВСМЯТКУ, В ВИДЕ, НАЛЕВО, НАПРАВО, НЕПОДАЛЕКУ, ПО-КАРСКИ, СБОКУ, СЛЕВА и т. п.; ср. ***дом неподалеку, крыша в виде купола, шашлык по-карски.***

РАЗЪЯСН - наречия, вводящие разъяснительную конструкцию: А ИМЕННО, В ТОМ ЧИСЛЕ, В ЧАСТНОСТИ, НАПРИМЕР, ТО ЕСТЬ и т. п.; ср. ***Обезьяны, например <в частности>, шимпанзе и гориллы, легко справляются с этой задачей.***

СУБ-МНОЖ - наречия, требующие "множественного" или "собирательного" субъекта (см. выше синтаксический признак ОБ-МНОЖ): ВМЕСТЕ, ВРАССЫПНУЮ, ПОРОЗНЬ, СООБЩА и т. п.; ср. ***Дети <все> бросились врассыпную, Петя, Катя и Маша бросились врассыпную, Петя с Машей бросились врассыпную, Толпа бросилась врассыпную*** при невозможности ****Ребенок бросился врассыпную.***

5. Признаки предлогов

ДИСКР - сложные предлоги, которые в сочинительной конструкции допускают повтор только последней своей части: В СВЯЗИ С, В СООТВЕТСТВИИ С, ПО ОТНОШЕНИЮ К, ПО НАПРАВЛЕНИЮ К, НЕСМОТРЯ НА и т. п.; ср. ***Несмотря на плохой урожай и на отсутствие внешней помощи, страна справилась с трудностями; по отношению к делегатам конференции и к их гостям.*** Простые предлоги повторяются в сочинительной конструкции целиком; ср. ***обратиться к делегатам конференции и к ее гостям.***

КОЛИЧ-ПР - предлоги, способные вводить количественные выражения: В1, НА1, ПО1; ср. ***высотой в три метра, больше на четыре человека, по трое больных.***

*Н-МЕСТ - предлоги типа БЛАГОДАРЯ, В КАЧЕСТВЕ, НАПОДО-

БИЕ, ПОДОБНО, ПОСРЕДСТВОМ, С ПОМОЩЬЮ и т. п., не вызывающие добавления эвфонического Н к местоименному существительному, т. е. мены типа ЕГО → НЕГО, ЕЕ → НЕЕ, ИХ → НИХ; ср. *благодаря ему* (*Н-МЕСТ), но *к нему* (не *Н-МЕСТ).

ОТРЫВ-НЕ - первообразные предлоги, способные вклиниваться между отрицательным глагольным элементом НЕ и основой соответствующего местоименного слова в комплексах типа НЕКОГО и НЕЧЕГО; ср. *Не к кому зайти, Не с кем поболтать, Не на что жить.*

ОТРЫВ-НИ - первообразные предлоги, способные вклиниваться между отрицательной частицей НИ и основой местоименных слов типа НИКТО, НИЧТО, НИКАКОЙ, НИЧЕЙ; ср. *ни к кому, ни в кого, ни с чем, ни с какими людьми, ни с чьими знакомыми*, но не **ни ради кого, *ни благодаря чему.*

ПЕРВООБР - первообразные предлоги типа В, ДЛЯ, ЗА, К, НА, НАД, О, ПО, ПОД, У, ЧЕРЕЗ и т. п., обладающие такими синтаксическими особенностями, как способность разрывать некоторые слова, некоторые композиты, вызывать появление эвфонического Н у местоимений; ср. *ни для кого, друг с другом, кое с кем, к нему* и т. п. Этот признак дополнителен к признаку *Н-МЕСТ.

ПУСТ - "пустой" сильноуправляемый предлог, подлежащий опущению при нормализации синтаксической структуры; свойство быть пустым зависит не от самого предлога, а от соответствующего управляющего слова и записывается в его модели управления; ср. *всматриваться в лицо, доходить до 20 градусов* (но *достигать 20 градусов*), *влиять на1 зрение* (но *отразиться на2 зрении*), *зависеть от глагола* (но *подчиняться глаголу*), *биться о борт корабля* (но *ударить в борт*), *подпадать под действие устава* (но *подвергаться действию*) и т. п. При других управляющих словах те же предлоги могут быть и непустыми; ср. *распространиться от Москвы до Владивостока, стоять под ковшом экскаватора* и т. п.

6. Признаки союзов

ВНУТР-ЗАП - составные подчинительные союзы (безусловные обороты) типа ПРИ УСЛОВИИ ЧТО, В СВЯЗИ С ТЕМ ЧТО, ПО МЕРЕ ТОГО КАК, ПОСЛЕ ТОГО КАК и т. п., внутри которых (в процессе синтаксического синтеза) должна быть поставлена запятая: *при условии, что; в связи с тем, что; по мере того, как.*

ПОДЧ - подчинительные союзы, ср. В ТО ВРЕМЯ КАК, ЕДВА, ЕСЛИ, КОГДА, ПОТОМУ ЧТО, С ТЕХ ПОР КАК, ТОГДА КАК, ЧЕМ, ЧТО1, ЧТОБЫ и т. п.; ср. ***Дверь распахнулась, едва он подошел.***

СОЧ - сочинительные союзы И, ИЛИ, А, НО, А ТАКЖЕ и т. п.

7. Признаки числительных

БОЛЬШ - "большие" числительные типа ПЯТЬ, ШЕСТЬ, ДЕСЯТЬ, ПЯТНАЛЦАТЬ, ДВАДЦАТЬ, ТРИСТА, ПЯТЬСОТ, НОЛЬ, вызывающие у хозяина-существительного в количественной конструкции форму множественного числа: ***семь книг*** (ср. ***две книги***).

ДВАДОД - составные числительные, оканчивающиеся на слово ОДИН и требующие при себе существительного в форме единственного числа; ср. ***(сто) двадцать один стол, тридцать одна машина, сорок одно замечание*** и т. п. Этот признак приписывается помещаемой в словаре фиктивной лексеме ЧИСЛ-ДВАДОД, которая используется при анализе чисел в цифровой форме.

МАЛ - "малые" числительные ОДИН, ПОЛТОРА, ДВА, ОБА, ТРИ, ЧЕТЫРЕ, вызывающие у хозяина-существительного в количественных конструкциях форму единственного числа; ср. ***четыре книги, полтора <два, три> мешка,*** но ***пять книг, шесть мешков.***

СОБИР - собирательные числительные типа ДВОЕ, ТРОЕ, ПЯТЕРО и т. п., которые в количественной конструкции способны сочетаться с названиями лиц мужского пола, словом ЛИЦО, названиями детенышей и некоторыми существительными pluralia tantum типа РЕБЯТА, САНИ, СУТКИ и т. п.; ср. ***двое <трое, пятеро> солдат, трое суток.***

Ц-ДРОБ - десятичная запись дробного числа меньше единицы, выступающая в роли хозяина в количественной конструкции; ср. ***0,87 объема раствора, 0,95 от общего количества.***

ЦИФ-1 - цепочка цифр, оканчивающаяся на 1 (в том числе цифра 1), но не на 11, и не содержащая внутри себя запятой, которая способна формировать количественную конструкцию, характерную для числительного ОДИН (см. выше признак ДВАДОД); ср. ***Пришел 21 человек.***

ЦИФР - любые числа, записанные цифрами и способные формировать разного рода количественные конструкции.

ЦИФР-БОЛЬШ - "большие" числительные, за исключением ДВА-

ДОД (см. признак БОЛЬШ), записанные цифрами; ср. *7 книг, 300 рублей*. Сюда входят также числа, соответствующие количественным существительным ТЫСЯЧА, МИЛЛИОН, МИЛЛИАРД.

ЦИФР-МАЛ - числительные, записанные цифрами и оканчивающиеся на 2, 3, 4 (но не на 12, 13, 14), которые сочетаются с существительными так же, как числительные с признаком МАЛ; ср. *2 стола, 3 книги, 32 бита, 1024 байта*.

Ц-СМЕШ - смешанная десятичная дробь, т.е. цепочка цифр, не начинающаяся с 0, внутри которой есть одна запятая и которая способна формировать количественную конструкцию, например, *7,85*.

ЧИСЛ-ДЦАТЬ - числительные ОДИННАДЦАТЬ, ДВЕНАДЦАТЬ,..., ДЕВЯТНАДЦАТЬ, а также порядковые прилагательные, обозначающие номера с 11-го по 19-й, способные употребляться в составе сложных имен чисел типа *сто одиннадцатый, двести девятнадцатый, триста одиннадцать* и т.п.

ЧИСЛ-Ц - числительные и порядковые прилагательные, обозначающие цифры и формирующие количественно-вспомогательную конструкцию типа *двадцать девять, сто первый*.

8. Признаки, приписываемые словам разных частей речи

ВВОД - наречия, наречные обороты и предлоги, способные формировать вводную конструкцию: ВЕРОЯТНО, ВИДИМО, ОЧЕВИДНО, ПО ВСЕЙ ВЕРОЯТНОСТИ, ПО-МОЕМУ, ПРИЗНАТЬСЯ, ПО1 и т.п.; ср. *Программа, очевидно <по всей видимости, по мнению руководителя>, содержит серьезные ошибки.*

ВВОД-i - глаголы и прилагательные, i-я (i = 1, 2, 3) валентность которых заполняется предложением с союзом ЧТО (ЧТОБЫ) и которые способны употребляться во вводной конструкции с союзом КАК: ВОДИТЬСЯ, ПОЛАГАТЬСЯ, СЛЕДОВАТЬ2; ИЗВЕСТНЫЙ, ПРИНЯТЫЙ ... (*Как известно, эти средства явно недостаточны*).

ВОПР - вопросительно-относительные местоимения, способные вводить косвенный вопрос или употребляться в сочинительных конструкциях с семантически однородными членами: ГДЕ, КАК, КОГДА, КОТОРЫЙ, КТО, КТО-НИБУДЬ, КТО-ЛИБО, КТО-ТО, КУДА, ОТКУДА, ЧЕЙ, ЧТО, ЧТО-ЛИБО, ЧТО-НИБУДЬ, ЧТО-ТО и т.п.: *Никто не знал, кого назначат руководителем работы <какими средствами институт будет располагать в будущем году, куда будут направлены выделенные институту средства>*;

Кто, куда и на какой срок командирован? По поводу последней конструкции заметим, что "нормальные" существительные, прилагательные и наречия, выполняющие такие разные синтаксические функции, как функции подлежащего [*кто*], обстоятельства направления [*куда*] и обстоятельства времени [*на какой срок*], сочиняться не могут; они требуют функциональной, т.е. синтаксической, однородности (подробнее о синтаксической и семантической однородности см. [Санников, 1989]).

ВРЕМ - существительные, наречия, предлоги и союзы, способные обозначать "точечную" локализацию ситуации во времени: ДЕНЬ, ЗАРЯ, КАНУН, ЧАС, ЭПОХА и т.п. (все они сочетаются с предлогом В1); НАЗАД, РАНО, СКОРО, ТОГДА; ДО, ПОСЛЕ, С1; КОГДА, ПОСЛЕ ТОГО КАК. Признак ВРЕМ охватывает все временные слова, в то время как соотносительный с ним признак ДЛИТ - лишь некоторые из них.

ДАТСУБ - существительные, прилагательные, наречия и наречные обороты, способные присоединять к себе свой первый актант в дательном падеже: БРАТ, ДРУГ, ЖЕНА, ОТЕЦ, СОСЕД и другие обозначения людей, являющиеся одновременно обозначениями отношений; ВЕСЕЛЫЙ, ГЛУБОКИЙ, МЕЛКИЙ, НУЖНЫЙ, ТЯЖЕЛЫЙ, ХОЛОДНЫЙ; ЛЕНЬ, ОХОТА, ЖАЛЬ, МОЖНО, НАДО, НЕЛЬЗЯ, НУЖНО, ПО КОЛЕНО, ПО ПОЯС и другие слова так называемой "категории состояния", обладающие тем же свойством; ср. *Он был мне другом <братом, отцом>, Мне здесь глубоко <мелко, холодно, по колено>.*

ДЕС - порядковые прилагательные и числительные, обозначающие десятки (номера второго разряда) и способные формировать количественно-вспомогательные конструкции - сложные имена чисел: ДЕСЯТЫЙ, ДВАДЦАТЫЙ, ТРИДЦАТЫЙ, СОРОКОВОЙ, ПЯТИДЕСЯТЫЙ,..., ДЕВЯНОСТЫЙ; ДЕСЯТЬ, ДВАДЦАТЬ, ТРИДЦАТЬ, СОРОК,..., ДЕВЯНОСТО; ср. *сто двадцатый <тридцатый>, сто сорок седьмой.*

ДЛИТ - уже упоминавшийся признак, приписываемый существительным и наречиям, способным употребляться в длительной конструкции типа *работать два часа <долго, недолго>.* Основной класс таких существительных - наименования отрезков времени типа МИНУТА, ЧАС, ДЕНЬ, ВЕЧЕР, НЕДЕЛЯ, МЕСЯЦ, ФЕВРАЛЬ, ГОД, ЭПОХА, ВРЕМЯ и т.п. Любопытно, что временные слова типа КАНУН, ЭРА признака ДЛИТ не имеют, поскольку не могут употребляться в конструкции типа **Он работал целый*

*канун, *Этот вид существовал целую эру.* Основной класс длительных наречий - наречия ДОЛГО и НЕДОЛГО.

ИСХОДН - наречия и предлоги со значением отправления из начальной точки: ОТКУДА, ОТСЮДА, ОТТУДА, ИЗ, ИЗ-ЗА, ИЗ-ПОД, ОТ, С1. Он вписывается в модели управления слов, имеющих валентность начальной точки, и используется для контроля семантического согласования между такими словами и возможными претендентами на заполнение этой валентности.

КАЧЕСТВ - качественные прилагательные и наречия, способные иметь степени сравнения и употребляться в синтаксических конструкциях с наречиями степени, наречиями со значением "точки зрения" типа МОРАЛЬНО, ПОЛИТИЧЕСКИ, ФИЗИЧЕСКИ, ЭКОНОМИЧЕСКИ и в ряде других синтаксических конструкций: БЕЛЫЙ, ВЕСЕЛЫЙ, ИНТЕРЕСНЫЙ, КРАТКИЙ, ЛЕГКИЙ, ПЛОХОЙ, ТЯЖЕЛЫЙ, ХОРОШИЙ, ВЕСЕЛО, ИНТЕРЕСНО, КРАТКО, ПЛОХО и т. п.; ср. *сильные духом, очень <чересчур> красивый, такой умный, самый интересный, какой хороший, лингвистически интересно, математически точно, морально устойчив.*

КВАНТ - кванторные прилагательные и наречия типа ВСЕ, ВСЯКИЙ, ЕДИНСТВЕННЫЙ, КАЖДЫЙ, НЕКОТОРЫЙ, ВЕЗДЕ, ВМЕСТЕ, ВСЕГДА, ВСЮДУ и т. п., цитируемые в разного рода обстоятельственных синтагмах.

КОЛИЧ - количественные существительные, прилагательные и наречия типа МИЛЛИОН, ДЕСЯТОК, ВСЕ, ОБА, МАЛО, МНОГО, НЕМНОГО, НЕСКОЛЬКО, ДОСТАТОЧНО, БОЛЕЕ, БОЛЬШЕ, МЕНЕЕ, МЕНЬШЕ и т. п., способные быть вершиной именных групп в разных количественных конструкциях; ср. *Сколько народу там было?*

КОМП - прилагательные и наречия со значением отличия, способные формировать сравнительную конструкцию, а также союзы, обладающие этим свойством: ДРУГОЙ, ИНОЙ, ПО-ИНОМУ, ПО-ДРУГОМУ, БОЛЕЕ, МЕНЕЕ, КАК, НЕЖЕЛИ, ЧЕМ и т. п.; ср. *Он предложил иной способ трансляции, чем принятый раньше; Он действует по-другому, нежели вы ожидали.*

КОНЕЧН - наречия и предлоги со значением направления к конечной точке: ТУДА, СЮДА, КУДА, ДОМОЙ, К, В1 (*во двор*), НА1 (*на улицу*), ЗА1 (*за дом*), ПОД1 (*под навес*), ПО НАПРАВЛЕНИЮ К и т. п. Этот признак, как и признак ИСХОДН, вписывается в модели управления и используется для проверки семантического согласования между управляющим словом и его потенциальным актантом.

ЛОКАТ - наречия и предлоги со значением чистой локализации: ЗДЕСЬ, ТУТ, ТАМ, ДОМА, В2 (*во дворе*), НА2 (*на улице*), ЗА2 (*за домом*), ПОД2 (*под навесом*) и т.п. Используется точно так же, как признаки ИСХОДН и КОНЕЧН.

МАЛ-СТ - прилагательные и наречия со значением "малого полюса" параметрической шкалы, не способные, в отличие от прилагательных и наречий со значением "большого полюса", сочетаться с вопросительной частицей ЛИ: КОРОТКИЙ, МАЛЫЙ, МЕЛКИЙ, НИЗКИЙ, РЕДКИЙ, УЗКИЙ, МАЛО, МЕЛКО, НИЗКО, РЕДКО, УЗКО и т.п.; ср. *Высока ли эта гора?, Глубока ли эта река?, Быстро ли ты бегаешь?, Часто ли вы навещаете друзей?* при невозможности **Низка ли эта гора?, *Мелка ли эта река?, *Медленно ли ты бегаешь?, *Редко ли вы навещаете друзей?* Это формальное противопоставление отражает тот содержательный факт, что только прилагательные со значением большого полюса способны выступать в русском языке в качестве имени соответствующей шкалы, т.е. шкалы высоты, глубины и т.д. Ср. *Как часто вы навещаете друзей? - Очень редко,* но не **Как редко вы навещаете друзей? - Очень часто.*

МЕСТ - местоименные существительные, прилагательные и наречия типа ЭТО, Я, КТО, ЧТО2, НЕКТО, КТО-ТО, КАЖДЫЙ, ЧЕЙ, КАКОЙ, КАК2, ЗДЕСЬ, ТУДА. Используется преимущественно для описания невозможных синтаксических конструкций; так, за некоторыми легко оговариваемыми исключениями (ср. *все это, ничего такого* и т.п.), существительное может быть вершиной определительной конструкции в случае, если оно не является местоимением; ср. *высокий мальчик,* но не **высокий он.*

НЕОПР - неопределенные местоименные существительные, прилагательные и наречия типа КТО-ЛИБО, КТО-НИБУДЬ, КТО-ТО, ЧТО-ЛИБО, ЧТО-НИБУДЬ, ЧТО-ТО, НЕКТО, НЕЧТО, КАКОЙ-ЛИБО, КАКОЙ-НИБУДЬ, КАКОЙ-ТО, КУДА-ЛИБО, КУДА-НИБУДЬ, КУДА-ТО и т.п., способные входить в "необычные" синтаксические конструкции; ср. *где-то на белом свете.*

ОБОБЩ - существительные, прилагательные и наречия типа АППАРАТУРА, ОБОРУДОВАНИЕ, ПУНКТ; РАЗНЫЙ, СЛЕДУЮЩИЙ; ВЕЗДЕ, ОТОВСЮДУ, ПОВСЮДУ. Все они выступают в качестве вводящих элементов в конструкциях типа *На завод поступило следующее оборудование: фрезерные и токарные станки, станки с программным управлением и запчасти к ним.*

ОГРАН - ограничительные частицы и наречия типа ДАЖЕ,

ЕДИНСТВЕННО, ЕЩЕ, И2, НЕ, НИ, ТАКЖЕ, ТОЖЕ, ТОЛЬКО, УЖЕ и т.п., способные подчиняться практически любым частям речи, ср. *Даже в восемь часов вечера Петр работает быстрее тебя, В восемь часов даже Петр работает быстрее, В восемь часов Петр даже работает быстрее, В восемь часов Петр работает даже быстрее.*

ОТН - относительные местоименнные существительные и наречия типа КОТОРЫЙ, ОТКУДА, ЧТО2 и т.п., способные формировать релятивные конструкции типа *Его перевели в отдел, в котором <где> начиналась разработка новой темы.*

ОТР - отрицательные существительные, прилагательные, наречия и частицы типа НИКТО, НИЧТО, НИКАКОЙ, НИ ОДИН, НИГДЕ, НИКОГДА, НИКУДА, НИ и т.п., требующие отрицания при глаголе: *Никто не приходил, Ничто ему не мешало работать, Никакой помощи оказано не было, Я никогда на это не соглашусь, Он не просил ни поблажек, ни снисхождения.*

ОТР-НЕ - субстантивные и адвербиальные комплексы НЕКОГО, НЕЧЕГО, НЕГДЕ, НЕЗАЧЕМ, НЕКОГДА, НЕКУДА, НЕОТКУДА, способные сочиняться друг с другом; ср. *Некуда, не к кому и незачем обращаться.* Субстантивные комплексы НЕКОГО и НЕЧЕГО обладают еще способностью разрываться первообразными предлогами: *Не с кем и не о чем говорить.*

ОТР-НИ - отрицательные существительные, прилагательные и наречия с приставкой НИ-, обладающие теми же двумя свойствами, что и только что рассмотренные местоименные комплексы; ср. *Никуда и ни к кому не обращайтесь.*

ОЦЕН - оценочные прилагательные и наречия, способные формировать ряд предикативных, комплетивных и обстоятельственных конструкций: ГЛУПЫЙ, ЛЕГКИЙ, ПЛОХОЙ, ТРУДНЫЙ, ТЯЖЕЛЫЙ, ХОРОШИЙ; ГРУБО, ЖЕСТОКО, НЕВАЖНО, ОТЛИЧНО, ПЛОХО, СУРОВО и т.п.; ср. *Звонить ему - глупая <пустая> затея, Дозвониться до него - трудное дело, Он выглядит хорошо, Мать относится к нему неважно <плохо>, Коллектив характеризует вас положительно <отрицательно>.*

ПОСТПОЗ - прилагательные, союзы и частицы, способные употребляться в постпозиции к своему синтаксическому хозяину, а также предлоги, способные находиться в постпозиции к своему слуге: ТОТ, ЭТОТ, РАДИ, СПУСТЯ, ПОТОМУ ЧТО, ТАК КАК, ЧТОБЫ, БЫ, ЖЕ, ЛИ, ТО и т.п.; ср. *Человек этот мне не понравился* (возможно и *Этот человек мне не понравился*), *Чего*

ради я должен стараться? (возможно и *Ради чего я должен стараться?*), *Он вернулся несколько лет спустя* (возможно и *Он вернулся спустя несколько лет*), *Мы постелили вам в соседней комнате, чтобы не беспокоить вас* (возможно и *Чтобы не беспокоить вас, мы постелили вам в соседней комнате*), *Я не промахнулся бы* (возможно и *Я бы не промахнулся*).

ПОСТПОЗ! - прилагательные и союзы, употребляющиеся только в постпозиции к своему синтаксическому хозяину: БЕЖ, БОРДО, БЛАГО2, ЗАТО, ПОТОМУ ЧТО, РАЗВЕ ЧТО, ТАК ЧТО, ТЕМ БОЛЕЕ ЧТО, ТОГДА КАК и т.п.; ср. *платье беж* (не **беж платье*), *Никого нет, так что некому жаловаться* (но не **Так что некому жаловаться, никого нет*).

ПРЕДВОПР - существительные, прилагательные и наречия, которые в функции предикатива (присвязочного члена) способны индуцировать у глагола-связки подлежащее в форме косвенного вопроса. Таковы слова ВОПРОС, ПРОБЛЕМА, НЕИЗВЕСТНЫЙ, ЯСНЫЙ, НЕЯСНЫЙ и ряд других. Ср. *Еще вопрос, поедет ли он туда; Неясно, какое решение он примет.* Мнемоническая структура этого признака такова: ПРЕД - "предикативное зависимое", или подлежащее; ВОПР - "(косвенный) вопрос".

Следует указать, что имеется целая серия признаков типа ПРЕД с аналогичным содержанием и аналогичной мнемонической структурой - ПРЕДИНФ, ПРЕДКОГДА, ПРЕДЧТО, ПРЕДЧТОБЫ. Они приписываются интенсиональным существительным, прилагательным и наречиям типа ЖЕЛАНИЕ, ЛОЖЬ, ОШИБКА, ПОЗОР, ПРАВДА, РАДОСТЬ, УДОВОЛЬСТВИЕ, ФАКТ, ЧЕСТЬ; ЛЕГКИЙ, НЕСОМНЕННЫЙ, ОЧЕВИДНЫЙ, ОПРЕДЕЛЕННЫЙ, ПОНЯТНЫЙ, СОМНИТЕЛЬНЫЙ, СПОРНЫЙ, ТРУДНЫЙ, ЯСНЫЙ; ЖАЛЬ, МОЖНО, НАДО, НЕЛЬЗЯ и т.п. Все такие слова суть интенсиональные предикаты, описывающие различные состояния сознания, эмоций и воли человека.

ПРЕДИНФ - то же самое, что ПРЕДВОПР, но в качестве подлежащего при связке выступает инфинитив: *Встретиться с вами было большим удовольствием <было приятно, было нельзя>.*

ПРЕДКОГДА - то же самое, но в качестве подлежащего при связке выступает придаточное предложение, вводимое союзами КОГДА1 или ЕСЛИ, ср. *Жаль, когда <если> принятые обязательства не выполняются.*

ПРЕДЧТО - то же самое, но в качестве подлежащего при связке выступает придаточное предложение, вводимое союзом ЧТО1, ср. *Является установленным фактом, что такие произ-*

водства экологически вредны; Сомнительно, что такие производства экологически безвредны.

ПРЕДЧТОБЫ - то же самое, но в качестве подлежащего при связке выступает придаточное предложение, вводимое союзом ЧТОБЫ, ср. *Сомнительно, чтобы такие производства были экологически безвредны.* Другие слова этого класса: НЕОБХОДИМОСТЬ, ПОТРЕБНОСТЬ; ЖЕЛАТЕЛЬНЫЙ, НЕВОЗМОЖНЫЙ, НЕОБХОДИМЫЙ, НАДО, НЕЛЬЗЯ, НУЖНО.

ПРЕД-ТО-ВОПР - существительные, прилагательные и наречия, обладающие следующим свойством: они могут быть присвязочной частью составного именного сказуемого, подлежащим при котором является придаточное вопросительное предложение, вводимое местоимением ТОЗ: ЗАГАДКА, ПРЕДМЕТ, ПРОБЛЕМА, ТЕМА; НЕСОМНЕННЫЙ, НЕЯСНЫЙ, РАЗУМНЫЙ, СОМНИТЕЛЬНЫЙ, ПОД ВОПРОСОМ и др.; ср. *То, куда он уехал, оставалось загадкой*

ПРЕД-ТО-КАК - то же самое, но в качестве подлежащего выступает придаточное предложение с союзом КАК; ср. *То, как сыграет чемпион, было загадкой <предметом многочисленных комментариев>.*

ПРЕД-ТО-ЧТО - то же самое, но в качестве подлежащего выступает придаточное предложение с союзом ЧТО1; ср. *То, что он согласится на это предложение, было несомненным <сомнительным>; То, что они не идут на поводу у общественного мнения, весьма похвально.*

ПРЕД-ТО-ЧТОБЫ - то же самое, но в качестве подлежащего выступает придаточное предложение с союзом ЧТОБЫ.

ПРЕПОЗ - подчинительные союзы и частицы, способные располагаться в препозиции к своему синтаксическому хозяину: ЕДВА, КОГДА, ТОЛЬКО2, ВЕДЬ, ДАЖЕ, ЕЩЕ, ЛИШЬ, НЕ, НИ, ТАКЖЕ, УЖЕ и т.п.; ср. *Едва он подошел, дверь распахнулась* (возможно и *Дверь распахнулась, едва он подошел*). Подчеркнем, что этот признак приписывается и таким союзам и частицам, как НЕ, ТОЛЬКО2, которые используются исключительно в препозиции (ср. признак ПРЕПОЗ! ниже).

ПРЕПОЗ! - прилагательные и союзы, способные располагаться только в препозиции к своему синтаксическому хозяину: КАКОЙ, ЧЕЙ, ТОЛЬКО2 и т.п.; *Какое право вы имеете делать ему замечания?, Чье это удостоверение?, Только он вышел, как все заговорили* (нельзя **Удостоверение чье это?, *Как все заговорили, только он вышел*).

ПРИБЛ - наречия и предлоги со значением неточной оценки количества, способные употребляться вместо именных групп в позициях дополнений, обстоятельств длительности и ряде других: БОЛЕЕ, БОЛЬШЕ, МЕНЕЕ, МЕНЬШЕ; СВЫШЕ, ДО, ЗА1, ОКОЛО, ОТ, ПОД1, С2 и т.п.; ср. *Их более 250 миллионов, Компьютер работал более <менее> двух часов, Техническая документация по Боингу весит свыше <около> 26 тонн, Пришло до двадцати человек.*

ПРИНИ - существительные и прилагательные типа ДЕНЬ, ДУША, КАПЛЯ, ЗВУК, МИНУТА, СЕКУНДА, ЧАС, ШАГ; ЕДИНЫЙ, МАЛЕЙШИЙ, ОДИН и т.п., способные присоединять непосредственно к себе частицу НИ; ср. *Вокруг ни души, Ни дня без строчки, Он не встретил у отца ни малейшего сочувствия.*

ПРИПРЕД! - наречия и предлоги, сочетающиеся только с глаголами или с существительными, обозначающими действия: ВРАЗВАЛКУ, ВСЛУХ, ОПТОМ, ПЕШКОМ, ПОСРЕДСТВОМ, С ПОМОЩЬЮ и т.п.; ср. *торговля <торговать> оптом, освобождаться <освобождение> от вируса посредством <с помощью> специальных программ.*

РАЗР - местоименные существительные, прилагательные и наречия типа НИКТО, НИЧТО, КОЕ-КТО, КОЕ-ЧТО, ДРУГ ДРУГА; НИКАКОЙ, НИЧЕЙ, КОЕ-КАКОЙ и т. п., способные разрываться первообразными предлогами; ср. *Он ни с кем не говорил об этом, Я хотел бы кое о чем поговорить с вами, Ни с чьими проектами они не знакомились.*

СИМ-2-1 - "симметричные" существительные, прилагательные, наречия и глаголы, требующие семантического согласования второго актанта с первым: ПАРАЛЛЕЛЬНОСТЬ, РАВЕНСТВО, СИММЕТРИЧНОСТЬ, ТОЖДЕСТВО, ЭКВИВАЛЕНТНОСТЬ; ПАРАЛЛЕЛЬНЫЙ, РАВНЫЙ, СИММЕТРИЧНЫЙ, ТОЖДЕСТВЕННЫЙ; ПАРАЛЛЕЛЬНО, СИММЕТРИЧНО; ЗНАКОМИТЬСЯ, ОБЩАТЬСЯ, РАВНЯТЬСЯ и т.п.; ср. *равенство треугольника ABC треугольнику DEF; прямая AC, параллельная прямой BE; Курс корабля проложен параллельно линии побережья, Новый директор познакомился с сотрудниками третьей лаборатории, Угол падения всегда равняется углу отражения.*

СИМ-3-2 - "симметричные" существительные и глаголы, требующие семантического согласования третьего актанта со вторым: ОТОЖДЕСТВЛЕНИЕ, СИНХРОНИЗАЦИЯ; ЗНАКОМИТЬ, ОТОЖДЕСТВЛЯТЬ, СИНХРОНИЗИРОВАТЬ, СОПОСТАВЛЯТЬ, СРАВНИВАТЬ и т.п.;

ср. *отождествление смелости с наглостью, сопоставлять "Союз" с "Аполлоном" по грузоподъемности.*

СИНТЕТ - прилагательные и наречия, имеющие синтетическую сравнительную степень: ВЫСОКИЙ, МАЛЫЙ, ПЛОХОЙ, ХОРОШИЙ; БЫСТРО, МЕДЛЕННО, ПЛОХО, ХОРОШО и т. п.; ср. *Он бегает быстрее всех в классе.*

СМЯГ - прилагательные и наречия, выступающие в форме аттенуативной сравнительной степени и способные употребляться в роли постпозитивного определения, вводимого частицей И2: ПОГЛУПЕЕ, ПОИНТЕРЕСНЕЕ, ПОЛУЧШЕ, ПОМОЛОЖЕ, ПОУМНЕЕ и т. п.; ср. *Я встречал девушек и покрасивее <поумнее>, Антенна расположена где-то повыше.*

*СОВ-2 - глаголы и существительные, которые не способны подчинять инфинитив совершенного вида на втором актантном месте (при допустимости несовершенного).: КОНЧАТЬ, НАЧИНАТЬ, ПЕРЕСТАВАТЬ, ПРОДОЛЖАТЬ и другие фазовые глаголы, разрозненные глаголы типа НАУЧИТЬСЯ, ОТВЫКАТЬ, ПЕРЕДУМАТЬ, ПРИВЫКАТЬ и т. п., а также ИСКУССТВО, НАУКА, ПРИВЫЧКА и т. п.; ср. невозможность **кончать <кончить> написать, *отвыкать <отвыкнуть> встать в семь часов, *искусство <наука> победить, *привычка встать*; ср. правильность *кончать писать, отвыкать вставать, искусство <наука> побеждать.*

*СОВ-3 - глаголы и существительные, не способные подчинять инфинитив совершенного вида на третьем актантном месте: ОТУЧАТЬ, ПРИУЧАТЬ, ПРИСПОСАБЛИВАТЬ, ОБУЧЕНИЕ и т. п.; ср. невозможность **отучать кого-л. встать в семь часов, *Первое занятие было посвящено обучению солдат пойти в строю* при правильности *отучать кого-л. вставать в семь часов, обучение кого-л. ходить в строю.*

СОГЛАКТ-3-2 - глаголы и существительные, требующие согласования третьего актанта со вторым по числу и роду: ДЕЛАТЬ, ДЕРЖАТЬ (*окна открытыми*), НАХОДИТЬ (*кого каким*), ПРИЗНАВАТЬ, СЧИТАТЬ; МНЕНИЕ, ПРИЗНАНИЕ, ОБЪЯВЛЕНИЕ и т. п.; ср. *Мы считаем <признаем> этот план* [ЕД, МУЖСК, 2-й актант] *вполне реальным* [ЕД, МУЖ, 3-й актант], *Мы считаем <признаем> эту программу* [ЕД, ЖЕНСК, 2-й актант] *реальной* [ЕД, ЖЕН, 3-й актант], *Мы считаем <признаем> эти планы* [МН, 2-й актант] *реальными* [МН, 3-й актант], *признание <объявление> Мадагаскара* [ЕД, МУЖСК, 2-й актант] *независимым* [ЕД, МУЖ, 3-й актант].

СОТН - порядковые прилагательные и числительные, обозначающие сотые номера и способные формировать количественно-вспомогательную конструкцию (сложные имена чисел): СОТЫЙ, ДВУХСОТЫЙ, ТРЕХСОТЫЙ,..., ДЕВЯТИСОТЫЙ, СТО, ДВЕСТИ, ТРИСТА,..., ДЕВЯТЬСОТ; ср. ***тысяча двухсотый, девятьсот семидесятый.***

ТОЗР - наречия со значением отношения, в котором рассматривается определенное свойство, способные сочетаться с прилагательными и наречиями, а также предлоги, входящие в состав фразем с указанным значением и синтаксической функцией: МАТЕМАТИЧЕСКИ, МОРАЛЬНО, ПО-СВОЕМУ, ПОЛИТИЧЕСКИ, ТЕОРЕТИЧЕСКИ, ФИЗИЧЕСКИ, В ПЛАНЕ, ПОД УГЛОМ ЗРЕНИЯ, С ТОЧКИ ЗРЕНИЯ и т. п. : ***политически правильное выступление, химически чистые вещества, физически безупречная теория, математически точно, Он абсолютно прав с точки зрения <под углом зрения, в плане> обычной морали.***

УКАЗ - указательные местоименные существительные, прилагательные и наречия типа ТО1, ЭТО, ТАКОЙ, ТОТ, ЭТОТ, ТАК, ТАМ, ТУДА и т. п., исключаемые на основании этого признака из ряда синтаксических конструкций.

ЭЛЕКТ - прилагательные и наречия типа БОЛЬШИЙ, ДРУГОЙ, ЛУЧШИЙ, ПОСЛЕДНИЙ, НЕСКОЛЬКО и т. п., способные быть вершиной элективной конструкции: ***больший <лучший> из них.***

6. 1. 2. 4. Зона семантических признаков

Аппарат синтаксических признаков дает основу для проверки синтаксического согласования между различными элементами конструкции. Однако одного этого аппарата мало, так как правильность синтаксических конструкций во многих случаях зависит и от семантической согласованности слов.

Так, актантные зависимые предикатного слова должны удовлетворять не только формальным, но и определенным семантическим требованиям, которые оно предъявляет к потенциальным заполнителям своих мест. Рассмотрим, например, слово СВОЙСТВО. Его первым (квазиагентивным) актантным зависимым должно быть слово, обозначающее предмет или вещество, ср. ***свойства каучука.*** В состав второго (1-го комплетивного) актантного зависимого должно входить слово со значением 'СВОЙСТВО', ср. ***свойство растяжимости.*** Положение в данном случае осложняется тем, что оба актантных зависимых присое-

диняются к ключевому слову в форме родительного падежа Если бы у нас не было средств для проверки семантической согласованности ключевого слова и его потенциальных слуг, на стадии СинтА для словосочетаний типа *свойство вязкости* из двух вырабатываемых гипотез - "квазиагент" и "1-компл" - нельзя было бы выбрать правильного СинтО (1-го комплетивного). Та же проблема возникает при анализе словосочетаний типа *прибивать что-л. гвоздями, пришивать что-л. суровыми нитками, прибивать что-л. молотком, пришивать что-л. специальной иглой*. Для правильного анализа подобных словосочетаний словам типа КАУЧУК, ВОДА, КИСЛОРОД и т.п. следует приписать семантический признак 'ВЕЩЕСТВО', а словам типа ВЯЗКОСТЬ, ЖЕСТКОСТЬ, ПРОЧНОСТЬ, РАСТЯЖИМОСТЬ, ТВЕРДОСТЬ и другим подобным - семантический признак 'СВОЙСТВО'. Одновременно при описании управления слов типа СВОЙСТВО необходимо указать, что от своего первого актанта они требуют признака 'ПРЕДМЕТ' или 'ВЕЩЕСТВО', а от второго - признака 'СВОЙСТВО'.

Другим участком синтаксической системы языка, где требуется проверка семантического согласования, являются сочинительные конструкции.

Вообще говоря, сочинительные группы в естественных языках строятся с соблюдением многих требований к свойствам однородных членов (см. [Санников, 1989]). У последних должны совпадать либо морфологические формы, либо синтаксические функции (ср. *Зачем и для кого это нужно*), либо коммуникативные (темо-рематические) функции (ср. *Пришел, но поздно*). К числу свойств, по которым в сочинительных группах происходит согласование однородных членов, относятся и семантические свойства. Нельзя, например, сочинять с помощью союза И обстоятельство (или дополнение) места и обстоятельство (или дополнение) времени; ср. неправильность (в сочинительной интерпретации) **Оборудование поступило на завод и в феврале*. Чтобы предотвратить возникновение гипотезы о возможности сочинения предложно-именных групп *на завод* и *в феврале*, достаточно приписать лексемам ЗАВОД и ФЕВРАЛЬ непересекающиеся наборы семантических признаков и потребовать, чтобы однородные (сочиняемые) члены предложения обладали хотя бы одним общим семантическим признаком.

Необходимость обращения к семантическим свойствам слов

может возникать и на этапе преобразования синтаксических структур в семантические. Мы, однако, не будем обсуждать этого вопроса подробно, полагая, что сказанного достаточно для обоснования необходимости использовать в модели анализа аппарат семантических признаков слов, или дескрипторов.

Дескрипторы в нынешней версии лингвистического процессора являются еще достаточно грубым инструментом описания семантических свойств слов, позволяющим определять лексические значения с точностью до очень больших классов; ср. названные выше семантические признаки 'ВЕЩЕСТВО', 'ПРЕДМЕТ' и 'СВОЙСТВО'. Впрочем, мы и не стремились к исчерпывающему описанию семантических свойств слов в терминах дескрипторов. Нам достаточно было того минимума дескрипторов, без которого синтаксические и семантические правила утратили бы эффективность. Этот минимум отыскивался, естественно, эмпирическим путем.

В нынешней версии лингвистического процессора для этапов синтаксического анализа и нормализации синтаксической структуры используется 18 семантических признаков. Условно можно говорить о собственных и несобственных семантических признаках.

Собственные семантические признаки лексемы - это часть ее значения. Они приписываются лексеме (слову или безусловному обороту) в зоне DES. Одной лексеме может быть приписано более одного признака, особенно в случае "укрупненных" значений (см. разд. 6.1.2.1). Так, лексеме КОНТАКТ, представляющей по крайней мере два разных значения этого слова, приписываются признаки 'ПРЕДМЕТ', 'ДЕЙСТВИЕ', 'СОСТОЯНИЕ', 'ФАКТ'.

Несобственные семантические признаки - это части лексических значений тех слов, с которыми сочетается данная лексема. Они приписываются лексеме не в зоне DES, а в зоне модели управления (см. след. разд.).

Семантические признаки приписываются в основном существительным и глаголам. Семантические свойства других частей речи достаточно полно описываются существующей, весьма разветвленной, номенклатурой синтаксических признаков.

Как ясно из сказанного, семантические признаки используются в правилах синтаксического анализа и нормализации синтаксических структур.

Для проверки семантического согласования используются два предиката - CORDES и CODES.

Требование CORDES дескрипторного согласования между управляющим словом X и управляемым словом Y считается выполненным, если Y имеет хотя бы один из дескрипторов, записанных в соответствующем столбце модели управления X, либо Y вообще не имеет дескрипторов, либо в соответствующем столбце модели управления X нет дескрипторов.

Требование CODES дескрипторного согласования слов X и Y, непосредственно входящих в сочинительную конструкцию или зависящих от сочиненных предлогов, считается выполненным, если у X и Y есть хотя бы один общий дескриптор, или хотя бы у одного из них вообще нет дескрипторов.

После этих предварительных замечаний можно перейти непосредственно к разбору дескрипторов.

'ВЕЛИЧИНА' - единицы измерения: ГОД, ДОЛЛАР, МЕТР и т.п. Дескриптор 'ВЕЛИЧИНА' очевидным образом дублирует синтаксический признак ИЗМЕР и вводится исключительно по содержательным соображениям. По этому дескриптору проверяется актантное семантическое согласование, для которого важен именно смысл, а не синтаксические свойства лексемы. Так, в комплетивной конструкции ***ускорение на 10 метров в секунду*** нет никакой синтаксической специфики, а важен лишь факт удовлетворения того семантического требования, которое лексема УСКОРЕНИЕ предъявляет к возможному претенденту на заполнение своего второго актантного места.

'ВЕЩЕСТВО' - названия веществ типа ВЕЩЕСТВО, ВОЗДУХ, ГАЗ, МАСЛО, МЕТАЛЛ, ПЛАЗМА и т.п.

'ВРЕМЯ' - временные слова типа БУДУЩЕЕ, ВРЕМЯ, ДЕСЯТИЛЕТИЕ, ПРОШЛОЕ и т.п.

'ДЕЙСТВИЕ' - слова, называющие действие, т.е. ситуацию, развивающуюся во времени и инициируемую активным субъектом (агенсом): АНАЛИЗ, ИЗМЕРЕНИЕ, РАБОТА, СОТРУДНИЧЕСТВО; АНАЛИЗИРОВАТЬ, ИЗМЕРЯТЬ и т.п.

'ИНФОРМАЦИЯ' - существительные типа ДАННЫЕ, ИНФОРМАЦИЯ, ОПЫТ, ПРОБЛЕМА, СМЫСЛ, СПИСОК, УРАВНЕНИЕ и т.п.

'КРИСТАЛЛ' - "прозрачные" существительные со значением части, порции или класса предметов: ВИД, КОМПОНЕНТ, ПОРЦИЯ, РОД, СЛОЙ, ЧАСТЬ и т.п. Они способны насыщать те же валентности слова, что и слова, обозначающие соответствующие

предметы. Как уже было сказано, в контексте "кристаллических" существительных дескрипторное согласование считается выполненным. Так, глагол КРАСИТЬ требует от своего третьего актанта семантического признака 'ВЕЩЕСТВО', ср. *покрасить дверь голубой краской*. Подыскивая в анализируемом предложении подходящее зависимое для КРАСИТЬ, мы не можем рассматривать в качестве допустимого кандидата на третье место, например, существительное ТЕОРИЯ, потому что оно не удовлетворяет требованию дескрипторного согласования. Действительно, конструкция **покрасить дверь теорией* была бы неправильной. Однако слово СЛОЙ может рассматриваться в таком качестве, потому что оно обладает признаком 'КРИСТАЛЛ'. Это позволяет опознать в качестве допустимой конструкцию типа *покрасить дверь двумя слоями голубой краски*.

'ЛИЦО' - название человека или организации: АГЕНТСТВО, АНАЛИТИК, ЛАБОРАТОРИЯ, ОПЕРАТОР, ПРАВИТЕЛЬСТВО, СПЕЦИАЛИСТ, ТЕХНОЛОГ и т. п.

'МЕХАНИЗМ' - названия приборов, механизмов, установок типа ДЕТЕКТОР, КОМПЬЮТЕР, ОСЦИЛЛЯТОР.

'ОТНОШЕНИЕ' - слова, обозначающие отношения: АНАЛОГИЯ, КОНТРАСТ, ОТНОШЕНИЕ, ПРИЧИНА, ПРОТИВОПОСТАВЛЕНИЕ, СООТВЕТСТВИЕ; СООТВЕТСТВОВАТЬ, ГРАНИЧИТЬ, ПРЕВОСХОДИТЬ и т. п.

'ПАРАМЕТР' - названия различных измеряемых параметров физических тел и процессов: ДАВЛЕНИЕ, КОНЦЕНТРАЦИЯ, МАССА, ОБЪЕМ, СКОРОСТЬ, ТЕМПЕРАТУРА, ЦЕНА и т. п. Семантический признак 'ПАРАМЕТР' в определенной мере дублирует синтаксический признак ПАРАМ. Основание для его введения - возможность компактнее сформулировать правила семантического согласования однородных членов в сочинительных цепочках типа *скорость и размеры ракеты*.

'ПРЕДМЕТ' - неодушевленный физический объект: ДЕТЕКТОР, ИОН, КОМПЬЮТЕР, ОРУДИЕ, ПЛАТФОРМА, СМЕСЬ, ЧИП и т. п.

'ПРОСТРАНСТВО' - название любого объекта, для которого его пространственная протяженность релевантна: АЭРОПОРТ, КУПЕ, МЕСТНОСТЬ, ОБЛАСТЬ, ПОВЕРХНОСТЬ, ПОЛОСТЬ, СКЛАД, СТРАНА и т. п. Этот признак используется в моделях управления таких слов, как ДОСТАВЛЯТЬ, ПРАВИТЕЛЬСТВО, СТАВИТЬ и других подобных, одно из актантных мест которых заполняется существительным с таким дескриптором: *правительство страны <штата>, доставлять груз на станцию* и т. п.

14 Заказ № 1538

'ПРОЦЕСС' - слова, называющие процессы, т.е. развивающиеся во времени ситуации с пассивным субъектом, в состоянии которого происходят изменения: БОЛЕЗНЬ, ПЛАВЛЕНИЕ, ПРОЦЕСС, РОСТ, ТАЯНИЕ, УБЫЛЬ, ФЛЮОРЕСЦЕНЦИЯ, ЭЛЕКТРОФОРЕЗ; ПЛАВИТЬСЯ, РАСТИ, ТАЯТЬ и т.п.

'СВОЙСТВО' - существительные, называющие более или менее постоянные свойства предметов, процессов, действий: БЕЗОПАСНОСТЬ, ВАЖНОСТЬ, ГИБКОСТЬ, КАЧЕСТВО, НАДЕЖНОСТЬ, СПОСОБНОСТЬ, УДОБСТВО и т.п.

'СОСТОЯНИЕ' - слова, называющие не развивающиеся во времени ситуации, время существования которых ограничено. Последним признаком они отличаются от свойств, ср. ВОЗМОЖНОСТЬ, ЗНАНИЕ, НУЖДА, ПОТРЕБНОСТЬ, ПРИСУТСТВИЕ, СОСТОЯНИЕ; ЗНАТЬ, ПРИСУТСТВОВАТЬ, ЛЕЖАТЬ и т.п.

'СФЕРА' - название области человеческой деятельности, включая названия всех наук: БИОЛОГИЯ, ЛИТЕРАТУРА, ОБЛАСТЬ, ПРОМЫШЛЕННОСТЬ, ТОКСИКОЛОГИЯ, ЭЛЕКТРОНИКА, ЭНЕРГЕТИКА и т.п.

'ФАКТ' - слова, обозначающие события: АВАРИЯ, КАТАСТРОФА, ПРОИСШЕСТВИЕ, ЯВЛЕНИЕ. Кроме того, дескриптор 'ФАКТ' дополнительно приписывается всем лексемам, имеющим дескрипторы 'ДЕЙСТВИЕ', 'ПРОЦЕСС', 'СОСТОЯНИЕ', 'СВОЙСТВО' или 'ОТНОШЕНИЕ'.

'ЭНЕРГИЯ' - существительные типа ИОН, ЛУЧ, НАПРЯЖЕНИЕ, ПОЛЕ, ТОК, ЭНЕРГИЯ. Этот дескриптор используется в моделях управления слов типа ДЕТЕКТОР, КОНЦЕНТРАЦИЯ, ПОГЛОЩЕНИЕ, СПЕКТР, ЭМИССИЯ и т.п.

Кроме перечисленных дескрипторов, обслуживающих этап анализа и преобразования текстов в любых прикладных системах лингвообработки, для подсистем общения, обслуживающих определенную предметную область (например, кадровую базу данных), вводятся дескрипторы этой области (например, 'СЛУЖАЩИЙ', 'ОТДЕЛ' и ряд других).

6.1.2.5. Зона моделей управления

В этой зоне описываются управляющие свойства предикатных слов, т.е. те требования, которые данный предикат предъявляет к потенциальным кандидатам на заполнение своих аргументных мест.

Речь идет лишь о канонических управляемых формах, т.е.

формах, которые с определенной степенью вероятности предсказываются самой данной лексемой. Вообще говоря, у лексемы в конкретном предложении могут быть и такие актантные управляемые зависимые, способ оформления которых диктуется не данной лексемой, а определенным синтаксическим контекстом. Они считаются неканоническими и непосредственно в модель управления лексемы не включаются.

Так, в русском языке в любой позиции, где возможна беспредложная именная группа в именительном, винительном или родительном падеже, возможна и аппроксимативно-количественная группа. Ср. ***Журнал опубликовал 40 статей о неевклидовых геометриях - Журнал опубликовал около <более> 40 статей - Журнал опубликовал от 40 до 50 статей - Журнал опубликовал до 40 статей - Журнал опубликовал более чем 40 статей.***

Эти синтаксические чередования абсолютно регулярны. Их можно естественно и просто описать как результат преобразования канонической управляемой формы в неканоническую под действием возмущающих контекстуальных синтаксических факторов. Следовательно, непосредственно на второе место в модели глагола ПУБЛИКОВАТЬ достаточно включить лишь каноническую управляемую форму ВИН. Неканонические управляемые формы в модель управления не вносятся.

На содержательном языке теории "Смысл ↔ Текст" (см., напр., [Мельчук, 1974]) модель управления лексемы имеет вид таблицы, состоящей из n столбцов - по числу актантов данной лексемы, т.е. участников обозначаемых ею ситуации действительности. В каждом столбце записывается информация о способах оформления данного актанта. Рассмотрим в качестве примера управляющие свойства существительного СРАВНЕНИЕ:

1	2	3	4
1.1 РОД, 'ЛИЦО' 1.2 ТВОР	2.1 РОД	3.1 СЗ	4.1 ПО1

Ср. ***сравнение учеными копии документа с оригиналом по всем существенным параметрам.***

На формальном языке лингвистического процессора всякая модель управления записывается в виде совокупности строк. Каждая строка вводится меткой вида Di.j, где i соответствует номеру столбца, а j - номеру строки в этом столбце, и

содержит совокупность термов или лексем t_1, t_2, ..., t_k. С учетом сказанного приведенная модель управления слова СРАВНЕНИЕ примет следующий вид:

D1.1:РОД,'ЛИЦО'·
D1.2:ТВОР
D2.1:РОД
D3.1:С3
D4.1:ПО1

Запятая, разделяющая термы в строке модели управления, имеет разные значения в зависимости от того, какие именно термы она разделяет. Если термы - не дескрипторы, то запись вида t_i, t_j обозначает конъюнкцию: потенциальное актантное зависимое данного предикатного слова должно обладать всеми указанными термами. Если же термы - дескрипторы, то такая запись обозначает дизъюнкцию: потенциальное актантное зависимое данного предикатного слова должно обладать хотя бы одним из них.

Строки модели управления могут содержать ссылки на следующие свойства актантов: 1) части речи; 2) конкретные лексемы; 3) синтаксические признаки; 4) дескрипторы; 5) морфологические характеристики; 6) символ НПУСТ (непустота управляемого предлога); 7) символ ВОПР, обозначающий способность управлять вопросительным предложением; 8) символ ЛИЧ, обозначающий личную форму глагола (бессоюзное управление целым предложением). Проиллюстрируем эти типы информации.

1. Части речи в модели управления. Глаголы со значением отношения и поведения типа ОТНОСИТЬСЯ, ОБРАЩАТЬСЯ, ПОСТУПАТЬ, ВЕСТИ (СЕБЯ), ДЕРЖАТЬСЯ и т.п. имеют на третьем (или втором) актантном месте помету ADV, ОЦЕН; ср. *относиться к кому-л. плохо, обращаться с кем-л. хорошо, поступать правильно, держаться высокомерно* и т.п. Глаголы так называемого среднего залога типа ВЕСИТЬ, ДЛИТЬСЯ, СТОИТЬ имеют на втором актантном месте помету ADV, КОЛИЧ или ADV, ДЛИТ, ср. *весить <стоить> много <немного>, длиться недолго.*

2. Конкретные лексемы в модели управления. Чаще всего в роли конкретных лексем выступают управляемые предлоги и союзы. Так, существительное РАБОТА управляет предлогами НАД, О2 и ПО1 на втором актантном месте: *работа над системой, работа о теории вероятностей, работа по лингвистике.* Глагол ПОДТВЕРЖДАТЬ и существительное УСЛОВИЕ управляют на

втором месте союзом ЧТО1, вводящим придаточное предложение: *подтверждать, что; при условии, что.*

3. Синтаксические признаки в модели управления. Наиболее часто в моделях управления различных слов (преимущественно глаголов и существительных) используются синтаксические признаки ИСХОДН, КОНЕЧН, ЛОКАТ. Словам ПОДНИМАТЬСЯ, ТЕЛЕГРАММА и другим подобным приписывается признак ИСХОДН на втором актантном месте и признак КОНЕЧН на третьем, ср. *подниматься со второго этажа на третий, телеграмма из Ленинграда в Москву.* Глаголам ЗАХОРАНИВАТЬ и РАСПРОСТРАНЯТЬ на третьем актантном месте приписывается признак ЛОКАТ, ср. *захоронить радиоактивные отходы в океане, распространить проект решения среди участников совещания.*

4. Дескрипторы в модели управления. В модели управления могут фигурировать любые дескрипторы, за исключением дескриптора 'КРИСТАЛЛ'. Так, на первом месте существительного ВМЕШАТЕЛЬСТВО записываются дескрипторы 'ЛИЦО', 'ФАКТ', 'ПАРАМЕТР', 'ЭНЕРГИЯ', а на втором - дескрипторы 'ДЕЙСТВИЕ' и 'ФАКТ', ср. *вмешательство властей, вмешательство в работу информационного центра.* Действие дескрипторов, как и действие синтаксических признаков, распространяется на все строки данного столбца. Как мы уже сказали, первое место существительного СРАВНЕНИЕ оформляется двумя способами, но дескрипторы записываются только в первой строке:

D1.1:ОД,'ЛИЦО'
D1.2:ТВОР

5. Морфологические характеристики в модели управления. Наиболее частая морфологическая характеристика - это падеж, что легко заметить на основании уже приведенных примеров. Другая именная характеристика - число. В частности, в модели управления симметричных предикатов вводится характеристика МН, описывающая ту реализацию управляющих свойств предиката, при которой два симметричных актанта представлены одной словоформой множественного числа: *Я познакомился с новым сотрудником - Мы познакомились, познакомить мальчика с девочкой - познакомить детей.* Из глагольных характеристик в модели управления может фигурировать ИНФ - для случаев типа *начинать делать что-л., решить сделать что-л.*

6. Символ НПУСТ в модели управления. Этот символ, как уже говорилось, обозначает "непустоту" предлога.

Пустыми называются такие управляемые предлоги, которые либо дублируют часть смысла управляющей лексемы, либо вовсе лишены самостоятельного лексического значения. Пустые предлоги в большинстве случаев весьма идиоматичны и автоматически предсказываются управляющей лексемой, ср. предлоги СЗ и ОТ при глаголах ПОЗДРАВЛЯТЬ (*кого-л. с чем-л.*) и ЗАВИСЕТЬ (*от чего-л.*).

Пустые предлоги при нормализации синтаксической структуры опускаются; ср.

зависеть $\xrightarrow{\text{1-компл}}$ *от* $\xrightarrow{\text{предл}}$ *X* ⇒ *зависеть* $\xrightarrow{\text{1-компл}}$ *X*.

Управляемый предлог, который обладает самостоятельным лексическим значением, называется непустым; ср. предлог ДЛЯ при слове АДАПТАЦИЯ или предлог ПО1 при слове СТРЕЛЬБА (*по чему-л.*). Непустые предлоги сохраняются в нормализованной структуре.

7. Символ ВОПР в модели управления. Этот символ фигурирует в моделях управления тех лексем, которые могут подчинять по какому-либо актантному отношению целую пропозицию, вводимую вопросительным словом. Таковы глаголы восприятия ВИДЕТЬ, ОЩУЩАТЬ, СЛЫШАТЬ и т. п. (*видеть <заметить>, куда он пошел <где он остановился>*); глаголы со значением интеллектуальных состояний и действий типа ЗНАТЬ, ОБЪЯСНЯТЬ, ПОНИМАТЬ, ПРОВЕРЯТЬ и ряд других (ср. *видеть <знать, понимать>, где <почему> программа делает ошибку*); глаголы со значением передачи и получения информации типа ГОВОРИТЬ, ОТВЕЧАТЬ, СПРАШИВАТЬ и т. п. (ср. *спрашивать <сообщать>, когда будут доставлены приборы*).

8. Символ ЛИЧ в модели управления. Этот символ фигурирует в моделях управления тех (относительно немногочисленных в русском языке) лексем, которые могут управлять придаточным дополнительным предложением, вводимым бессоюзно. К их числу относятся, например, глаголы ВИДЕТЬ и ГОВОРИТЬ; ср. *Он говорит <Я вижу>, вы устали.*

6. 1. 3. Операционная информация в словарной статье КС

Как мы уже говорили, в словарную статью КС может быть включено два типа операционной информации: ссылка на трафаретное правило (в виде имени правила) и словарное правило (целиком). Поскольку сами правила были подробно описаны в

предшествующих разделах, мы приведем здесь лишь представительные примеры правил следующих трех типов: 1) синтаксического анализа; 2) нормализации; 3) семантизации.

6.1.3.1. Правила синтаксического анализа

Трафаретное правило на существительные, обладающие семантическими признаками 'ДЕЙСТВИЕ' или 'ПРОЦЕСС': анализ атрибутивных конструкций типа ***обработка деталей давлением.***

```
REG:АТРИБ.15
LOC:RALFA
RALFA:'ДЕЙСТВИЕ'/'ПРОЦЕСС'
N:Ø1
CHECK
1.1 =(X,RALFA,ТВОР)&L-EQU(X,Y,5,S,RALFA)
1.2 ^ININT(Y,X,ЛИЧ,ДЕЕПР,ИНФ,ПОДЧ)
DO
1 SVUZOT:(Y,X,АТРИБ)
```

Словарное правило на слово ТАКОЙ: синтаксический анализ комплетивных конструкций типа ***Рассмотрим такой треугольник, что две его стороны равны.***

```
REG:1-КОМПЛ.OO
NONPR:1
N:Ø1
CHECK
1.1 R-LEXR(X,Y,4,ЧТО1,ЧТОБН,КАК1)&PNLEFT(Y,ЗПТ)
3.1 I-DEP1(X,*,Y,3,ОПРЕД)
3.2 ^LEXR(Y,ЧТОБН)/DOM-LEXR(X,*,ОГРАНИЧ,НЕ)/^DOM(Y,*,
    ИНФ-СОЮЗН)
3.3 ^LEXR(Y,КАК1)/DOM-EQU(Y,*,ПОДЧ-СОЮЗН,ИМ)
DO
1 SVUZOT:(X,Y,1-КОМПЛ)
```

6.1.3.2. Правила нормализации

Трафаретное правило восстановления опущенного существительного по субстантивированному причастию - на глаголы, причастия которых допускают субстантивацию: ВОЗГЛАВЛЯТЬ, ЗАНИМАТЬСЯ, ЗАРАБАТЫВАТЬ, РАБОТАТЬ и т. п.

```
REG:NORMAL2.Ø2
LOC:RR
```

```
RR:ПРЕДИК/1-КОМПЛ/КВАЗИАГЕНТ
N:Ø1
CHECK
1.1 = (X,ПРИЧ)&#(X,СТРАД)&DEP(X,*,RR),&VAL(1,X,'СЛУЖАЩИЙ')
DO
1 DOBRUZ:Z(СЛУЖАЩИЙ)
2 PERUZHAR:X(NMB)-Z
3 PERUZPREDNOM:Z(X)
4 SVUZOT:(Z,X,ОПРЕД)
5 IZSLOT:(X,RR)-(Z,RR)
```

Примеры: ***Какие должности занимают работающие*** [X] ***в Чикаго?*** → ***Какие должности занимают служащие*** [Z], ***работающие*** [X] ***в Чикаго? Подсчитать количество зарабатывающих 300 долларов*** → ***Подсчитать количество служащих, зарабатывающих 300 долларов.***

Словарное правило восстановления опущенного существительного в сравнительной конструкции с союзом КАК1 - на этот союз.

```
REG:NORMAL.ØØ
N:Ø1
CHECK
1.1 DOM-LEXR(X,Y,СРАВН-СОЮЗН,СРЕДНИЙ,МАКСИМАЛЬНЫЙ,
     НАИБОЛЬШИЙ,МИНИМАЛЬНЫЙ,НАИМЕНЬШИЙ)&DOM(Y,U,*)
1.2 DEP-EQU(X,Z,СРАВНИТ,S)/DEP(X,Z1,СРАВНИТ)&DEP-EQU
     (Z1,Z,*,S)/DEP(X,Z2,СРАВНИТ)&DEP(Z2,Z1,ОПРЕД)
     &DEP-EQU(Z1,Z,*,S)
DO
1 DOBRUZ:W(ФИКТ-ЛЕКС)
2 PERLEKRUZ:Z(W)
3 PERUZHAR:Z(NMB)-W
4 PERUZSLEDNOM:W(Y)
5 IZSLOT:(Y,*)-(W,СРАВН-СОЮЗН)
6 PERUZOT:(Y,U,*)-(W,U,КВАЗИАГЕНТ)
7 SVUZOT:(W,Y,ОПРЕД)
```

Примеры: ***зарплата*** [Z], ***как*** [X] ***средняя*** [Y] (***в***) ***отделе*** [U] ***20; такая*** [Z1] ***же зарплата*** [Z], ***как*** [X] ***средняя*** [Y] (***в***) ***отделе*** [U] ***20; та*** [Z2] ***же самая*** [Z1] ***зарплата*** [Z], ***как*** [X] ***средняя*** [Y] (***в***) ***отделе*** [U].

6.1.3.3. Правила семантизации

Трафаретное правило замены русских слов МАКСИМАЛЬНЫЙ, МИНИМАЛЬНЫЙ, СРЕДНИЙ в определительном контексте соответствующими словами семантического языка.

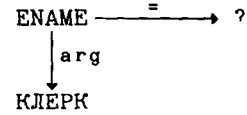
```
REG:SEMANT1.1Ø
TAKE:1
N:Ø1
CHECK
1.1 DEP-EQUN(X,U,ОПРЕД,FNC,OPR)
DO
1 ZAMAUZ:X(LA)
2 IZSLOT:(U,*)-(X,*)
3 SVUZOT:(X,U,ARG)
```

Комментарий. Это правило заменяет указанные выше русские слова их семантическими коррелятами - словами MAX, MIN, AVG (average). Одновременно они превращаются в вершину соответствующего поддерева, а их бывший хозяин - слово типа ЗАРПЛАТА - становится их слугой по аргументному отношению. Иными словами, происходит преобразование типа *максимальная зарплата* ⟶ MAX *зарплата*.

Трафаретное правило замены русских слов УКАЗАТЬ, ПЕРЕЧИСЛИТЬ, НАЗВАТЬ в формах инфинитива или повелительного наклонения вопросительной константой '?' в различных лексических контекстах.

```
REG:SEMANT1.11
TAKE:1
LOC:R
R:1-КОМПЛ/ПРЕДИК
CHECK
1.1 DOM(X,U,R)
N:Ø1
CHECK
1.1 =(U,'ЧЕЛОВЕК')
DO
1 ZAMAUZ:(X,ENAME)
2 IZOT:(X,U,*)-(X,U,ARG)
3 DOBAUZ:W('?')
4 SVUZOT:(X,W,EQU)
5 PERUZSLEDNOM:W(X)
```

Комментарий. Фраза ***Указать клерков*** преобразуется этим правилом в структуру вида



Константа '?' переносится на следующее после слова ENAME место по чисто техническим причинам - для удобства рассмотрения таких деревьев и их контроля со стороны человека на стадии экспериментальной отладки системы.

Другие подправила этого правила проводят аналогичные преобразования для случаев типа ***указать отделы, указать город, указать штат, указать зарплату*** и т. п.

6.1.4. Образцы словарных статей КС

Ниже приводятся словарные статьи трех слов - ДАВЛЕНИЕ, МАКСИМАЛЬНЫЙ и РАБОТАТЬ.

```
Ø3918                          ДАВЛЕНИЕ
     POR S
     SYNT СРЕДН,ЕД',ПАРАМ,ХАРАКТТВОР
     DES 'ДЕЙСТВИЕ','ПРОЦЕСС','ФАКТ'
     D1 1 РОД
     D2 1 НА1
     D3 1 В1,'ВЕЛИЧИНА'
     D3 2 ИМ,ИЗМЕР
TRAF АТРИБ 15

ØØ35Ø                          МАКСИМАЛЬНЫЙ
     POR A
     SYNT КАЧЕСТВ,КВАНТЪЕД-СРЕД,ОЦЕН-РЕЛЯТ,ЭЛЕКТ
TRAF РЕЛЯТ 11
TRAF РЕЛЯТ 13
TRAF ДЛИТЕЛЬН 14
TRAF ОБСТ 22
TRAF SEMANT1 1Ø
LA MAX
TRAF EXPANS 73
LR МАКСИМАЛЬНО

Ø15Ø7                          РАБОТАТЬ
     POR V
     SYNT НЕСОВ',КОНСТРУКТ,СКОПР
```

```
      DES 'ДЕЙСТВИЕ','ФАКТ'
      D2 1 ТВОР,'ДОЛЖНОСТЬ'
      D2 2 В_КАЧЕСТВЕ
      D3 1 В2,'ОТДЕЛ'
      D3 2 ADV,ЛОКАТ
      D4 1 НАД
REG 1-КОМПЛ 00
N Ø1
CHECK
1 1 M-LEXR(X,Y,1Ø,B2,HA2)&R-LEXR(Y,Z,Ø,ДОЛЖНОСТЬ)
3 1 DOM(Y,Z,ПРЕДЛ)&DOM-EQU(Z,*,1-КОМПЛ,'ДОЛЖН')
DO
1 SVUZOT (X,Y,1-КОМПЛ)
TRAF ОБСТ 21
TRAF 1-КОМПЛ 20
TRAF 1-КОМПЛ 21
TRAF СУБ-КОПР 10
TRAF NORMAL1 15
TRAF NORMAL2 2
TRAF SEMANT1 23
TRAF SEMANT1 24
TRAF EXPANS 64
LR РАБОТА
TRAF EXPANS 93
LR НАД
TRAF SYNTHES1 13
```

6.2. Семантический словарь

6.2.1. Типы информации в семантическом словаре

В этом разделе будет описан второй тип словарей, используемых в ЛП, - семантический словарь (СС). Как по количеству словарных статей, так и по объему информации, содержащейся в каждой словарной статье, он значительно уступает комбинаторному словарю, описанному в предыдущем разделе. Это обстоятельство не случайно.

Во-первых, как отмечалось в разд. 6.1, морфологический и синтаксический компоненты ЛП не ориентированы на какую-либо конкретную предметную область и могут быть использованы в широком круге информационных систем, включая системы машинного перевода. Поэтому КС, на который существенно опи-

рается синтаксический компонент ЛП, должен включать большое число слов, которые заведомо не релевантны для задачи общения с БД.

Во-вторых, и это особенно важно, в отличие от КС объектом СС не являются слова естественного языка. Этот словарь описывает элементарные единицы семантического языка, из которых строятся семантические структуры запросов, - константы, функции, операторы и переменные. Из определения СемС (см. разд. 5.3.4) и из примеров, которые приводились в предыдущей главе, легко видеть, что СемС как формальный объект устроена весьма просто. Она состоит из одного или нескольких элементарных предикатных поддеревьев, каждое из которых, в свою очередь, построено по простым правилам из элементарных единиц семантического языка. Между тем, чем более стандартизованы и единообразны выражения, построенные из единиц данного языка, тем менее индивидуально поведение каждой из этих единиц и, следовательно, тем меньший объем информации надо приписать ей в словаре.

Принципы описания словарных единиц и общий формат словарной статьи в семантическом словаре в целом не отличаются от комбинаторного словаря. Информация, помещаемая в словарную статью, также делится на два типа - классификационную и операционную. Иным, более редуцированным, является лишь состав зон словарной статьи.

Классификационная информация представлена тремя разновидностями. Это аналоги частей речи, синтаксических и семантических признаков.

Часть речи задает собственный синтаксический класс данной единицы. Набор этих классов невелик - CONST (константа), FNC (функция), OPR (оператор) и VAR (переменная). Каждая единица имеет признак из этого списка и притом только один.

Синтаксический признак позволяет указать, в случае функциональной единицы, какой элемент может служить ее аргументом. Поскольку аргументом функции, по определению, может быть только переменная, то таких признаков столько, сколько переменных.

ARG-EM (аргумент функции - переменная EM): этот признак приписывается функциям, характеризующим свойства служащих, - ENAME (фамилия), JOB (должность), SAL (зарплата) и др.

ARG-DP (аргумент функции - переменная DP): присваивается функциям, задающим признаки отделов - DEPTNO (номер отдела), DNAME (название отдела), LOC (местонахождение отдела), MGR (фамилия менеджера) и др.

ARG-CT (аргумент функции - переменная CT): этим признаком обладают функции, описывающие города - CITY (название города) и STATE (название штата, в котором расположен данный город).

ARG-ST (аргумент функции - переменная ST): этот признак характеризует функции, приписанные штатам - SNAME (название штата).

Синтаксические признаки предназначены для описания сочетаемостных свойств функций. Однако приписываются они не только функциям, но и соответствующим переменным (признак ARG-EM - переменной EM, признак ARG-DP - переменной DP и т.д.). Это делается для того, чтобы проверка согласования функции с переменной сводилась к обнаружению у согласуемых элементов одного и того же признака. Функция может иметь более одного синтаксического признака - ср., например, функцию DEPTNO, которая сочетается как с переменной EM, так и с переменной DP.

Семантические признаки приписываются константам, функциям и операторам и задают их семантический класс. Используются следующие семантические признаки:

'ИМ-СЛУЖ' (имя служащего) - приписывается константам типа 'Смит' и функциям, принимающим значения на этом множестве констант (ENAME, MGR);

'ИМ-ОТД' (название отдела) - приписывается константам типа 'бухгалтерия' и соответствующим функциям (DNAME);

'ИМ-ГОР' (название города) - приписывается константам типа 'Чикаго' и соответствующим функциям (LOC, CITY);

'ИМ-ШТАТ' (название штата) - приписывается константам типа 'Калифорния' и соответствующим функциям (STATE, SNAME);

'ДОЛЖН' (должность) - приписывается константам типа 'клерк' и соответствующим функциям (JOB);

'ЧИСЛ' (число) - приписывается функциям и операторам, принимающим числовые значения (SAL, DEPTNO, MAX, MIN, AVG, COUNT).

Если синтаксические признаки описывают согласование чле-

нов аргументного отношения, то семантические признаки служат для проверки согласования членов предикатных отношений (таких, как =, ≠, <, > и т.п.).

Как правило, слово может иметь только один семантический признак. Это вполне естественно: либо оно служит названием города, либо обозначает число, но никак не то и другое вместе. Однако у этого правила есть исключения. В семантическом языке есть своего рода местоимения - константы, способные вступать в предикатные отношения с функциями любого семантического типа. Таких местоимений два - вопросительная константа '?' и константа с общеквантворным значением 'all', соответствующая слову КАЖДЫЙ. Примеры СемС с этими константами в разных контекстах мы приводили в разд. 5.4.2.1 (см. (17)-(20)).

Операционная зона словарной статьи СС содержит отсылки к правилам семантизации или оптимизации СемС.

Между классификационной и операционной информацией в семантическом словаре имеется следующее принципиальное различие. Операционная информация целиком определяется теми правилами, которые составлены разработчиком, и в этом смысле зависит от его произвола. Классификационная информация в каждой словарной статье, как и сам словник семантического языка, не зависит от разработчика, а задается ему извне, со стороны БД.

Мы уже отмечали в главе 5, что семантическая структура в ЛП обращена в сторону предметной области, задаваемой реляционными таблицами БД. Семантические элементы, из которых строятся СемС и которые образуют семантический словарь, непосредственно заимствуются из двух источников, каждый из которых лежит вне ЛП. Большая часть из них извлекается из реляционных таблиц БД. Названия таблиц превращаются в переменные, атрибуты таблиц - в функции, а значения атрибутов - в константы. Остальные семантические элементы - операторы - в таблицах БД не представлены, поскольку предназначены для выполнения логических и арифметических операций над данными таблиц. Эти элементы рекрутируются из числа операторов языка SQL.

Этот принцип формирования семантического словаря естественным образом задает классификационную информацию в его словарных статьях. Часть речи семантического элемента опре-

деляется тем, из какого источника он заимствован. Синтаксические признаки функций обусловлены тем, в каких таблицах выступают соответствующие им атрибуты. Семантические признаки диктуются областью значений атрибутов и характером действия, осуществляемого оператором.

Приведем теперь несколько примеров словарных статей семантического словаря разного типа.

6.2.2. Образцы словарных статей семантического словаря

```
Ø0228                                        'КЛЕРК'
          POR:CONST
          DES:'ИМ-ДОЛЖН'
Ø0248                                        'КОММЕРЧЕСКИЙ'
          POR:CONST
          DES:'ИМ-ОТД'
Ø025Ø                                        'ДЖОУНЗ'
          POR:CONST
          DES:'ИМ-СЛУЖ'
Ø0249                                        'НЬЮ-ЙОРК'
          POR:CONST
          DES:'ИМ-ГОР','ИМ-ШТАТ'
Ø027Ø                                        'ALL'
          POR:CONST
          DES:'ДОЛЖН','ИМ-СЛУЖ','ИМ-ОТД','ИМ-ГОР',
              'ИМ-ШТАТ','ЧИСЛ'
TRAF:SEMANT2.12
Ø0Ø08                                        DNAME
          POR:FNC
          SYNT:ARG-DP
          DES:'ИМ-ОТД'
Ø0Ø06                                        SAL
          POR:FNC
          SYNT:ARG-EM
          DES:'ЧИСЛ'
Ø0Ø07                                        DEPTNO
          POR:FNC
          SYNT:ARG-EM,ARG-DP
          DES:'ЧИСЛ'
```

```
ØØØ2Ø                                          MAX
            POR:OPR
            DES:'ЧИСЛ'
TRAF;SEMANT2.6Ø
TRAF:SEMANT.1Ø
TRAF:SEMANT.11
TRAF:SEMANT.12
TRAF:SEMANT.13
TRAF:SEMANT.14
ØØØØ1                                          EM
            POR:VAR
            SYNT:ARG-EM
TRAF:SEMANT2.Ø8
TRAF:SEMANT2.Ø9
TRAF:SEMANT2.1Ø
TRAF:SEMANT2.18
TRAF:SEMANT2.19
TRAF:SEMANT2.4Ø
TRAF:SEMANT.Ø7
TRAF:SEMANT.Ø8
```

Глава 7
ЭКСПЕРИМЕНТЫ ПО СЕМАНТИЧЕСКОМУ АНАЛИЗУ И МАШИННОМУ ПЕРЕВОДУ

Завершив разбор всех лингвистических компонентов и алгоритмических модулей ЛП, мы дадим в этой главе краткий обзор той экспериментальной работы, которую мы с ним провели. Хронологически и по существу она распадается на три главные части: 1) эксперименты по МП с английского языка на русский (система ЭТАП-2); 2) эксперименты по семантическому анализу, или перевод запросов к реляционной БД с естественного языка на SQL; 3) эксперименты по МП с русского языка на английский и обратно на базе расширенной системы ЭТАП-2 (система ЭТАП-3).

Первые эксперименты по МП с английского языка на русский были проведены нами еще в 1984-1985 гг. в институте "Информэлектро" на ЭВМ ЕС-1033. Тогда же, в связи с переходом нашей лингво-математической группы в ИППИ АН СССР, они были надолго законсервированы. Лишь спустя два с половиной года система ЭТАП-2 была возрождена на новой технической базе и серьезно расширена и усовершенствована. Существо всех изменений, которые претерпело лингвистическое обеспечение системы, было рассмотрено в предшествующих главах. Естественно, что усовершенствования в лингвистическом и логико-алгоритмическом обеспечении благотворно сказались как на качестве перевода, так и на скорости его получения. Однако отличие нового экспериментального материала от того, который был приведен в нашей монографии [Апресян и др., 1989], не настолько велико, чтобы специально его излагать и в этой книге. Поэтому ниже мы рассмотрим в основном результаты двух последних серий экспериментов - по семантическому анализу и по русско-английскому МП.

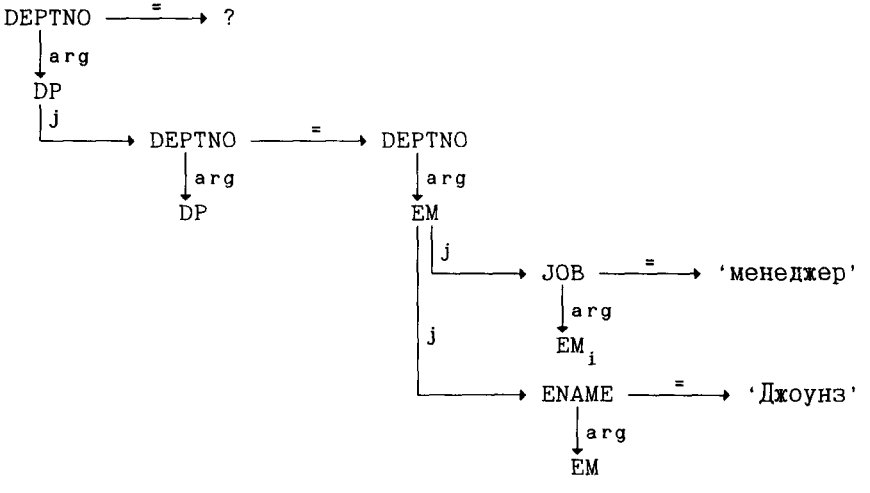
7.1. Эксперименты по семантическому анализу

После подробного описания семантического анализа, данного в гл. 5, будет достаточно привести здесь некоторые результаты экспериментов, которые мы проводили с ЛП, присо-

единенным к СУБД ORACLE. В экспериментах запросы к БД предлагались на естественном языке. Лингвистический процессор переводил их на язык SQL и передавал СУБД для исполнения. Если СУБД воспринимала полученное SQL-представление и давала на него ожидаемый ответ, то это было критерием того, что ЛП работает адекватно.

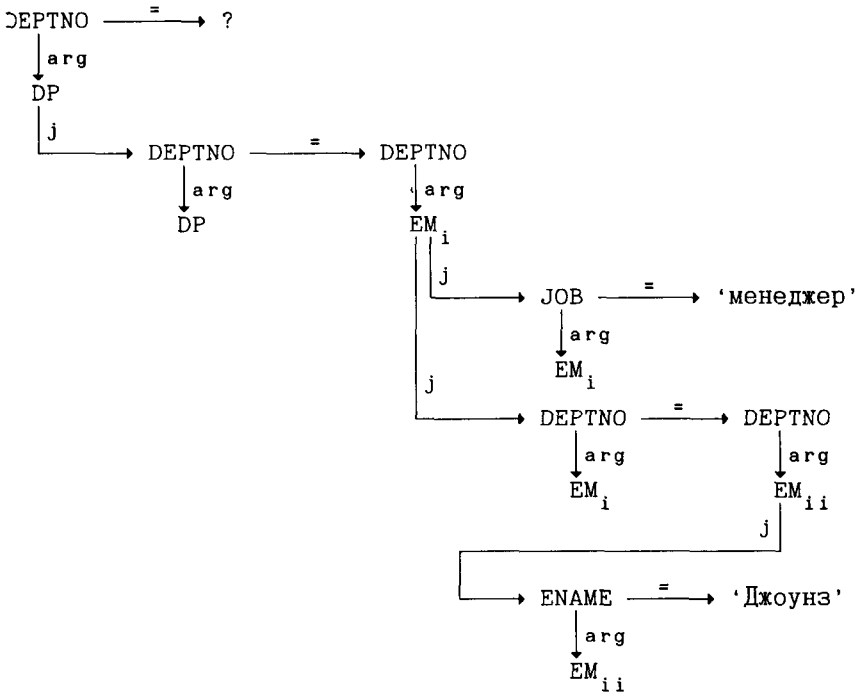
Ниже будут указаны сами запросы, полученные для них семантические структуры и SQL-представления. Запросы для иллюстрации подбирались таким образом, чтобы по ним одним заинтересованный читатель мог составить себе представление о мощности ЛП и об объеме естественно-языковых знаний, которыми он обладает. Например, запросы (1) и (2), различающиеся лишь синтаксической функцией одного имени, убеждают, что неучет сведений подобного рода в общем случае не позволяет рассчитывать на правильный результат. Запросы (3) и (4) показывают, что ЛП способен правильно уловить различие между местоимениями ЕГО и СВОЙ в условиях, когда эти местоимения различаются синтаксической функцией имени, претендующего на роль антецедента. В запросе (5) можно обратить внимание на обработку сочинительных конструкций: союз И соответствует в одних случаях конъюнкции, а в других - дизъюнкции. Наконец, запрос (6) иллюстрирует возможности, которыми обладает ЛП для анализа операторных слов типа МАКСИМАЛЬНЫЙ.

(1) В каком отделе работает менеджер Джоунз?



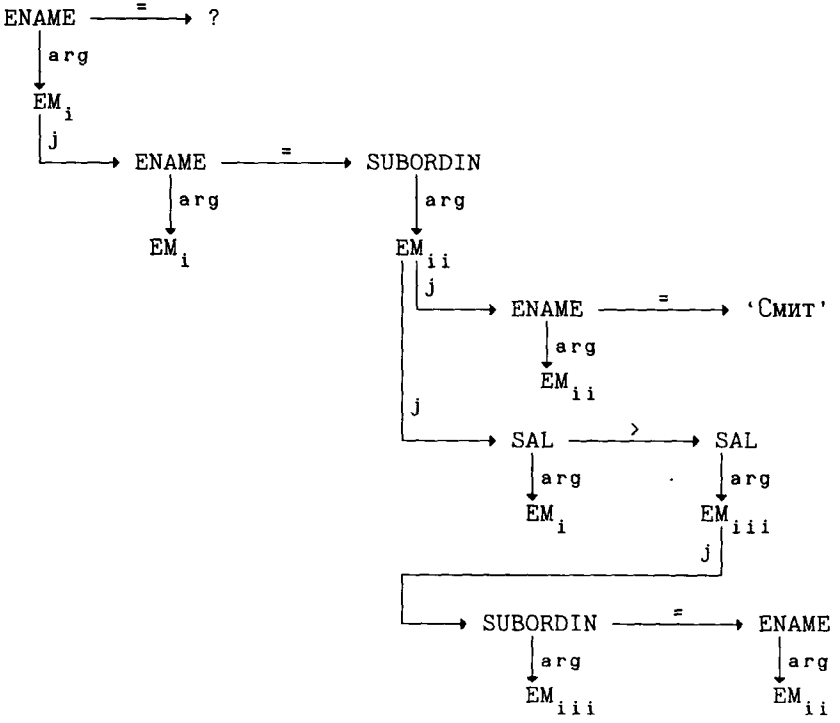
```
SELECT  DEPTNO
FROM    DP A
WHERE   (A.DEPTNO IN
                (SELECT  DEPTNO
                 FROM    EM B
                 WHERE   (B.JOB = 'менеджер' AND B.ENAME =
                                                 'Джоунз')))
TIME:ØØ.18
```

2) В каком отделе работает менеджер Джоунза?




```
SELECT  DEPTNO
FROM    DP A
WHERE   (A.DEPTNO IN
                (SELECT  DEPTNO
                 FROM    EM B
                 WHERE   (B DEPTNO IN
                                 (SELECT  DEPTNO
                                  FROM    EM D
                                  WHERE   (D.ENAME =
                                          'Джоунз')))
                    AND   (B.JOB = 'менеджер')))
TIME:ØØ.17
```

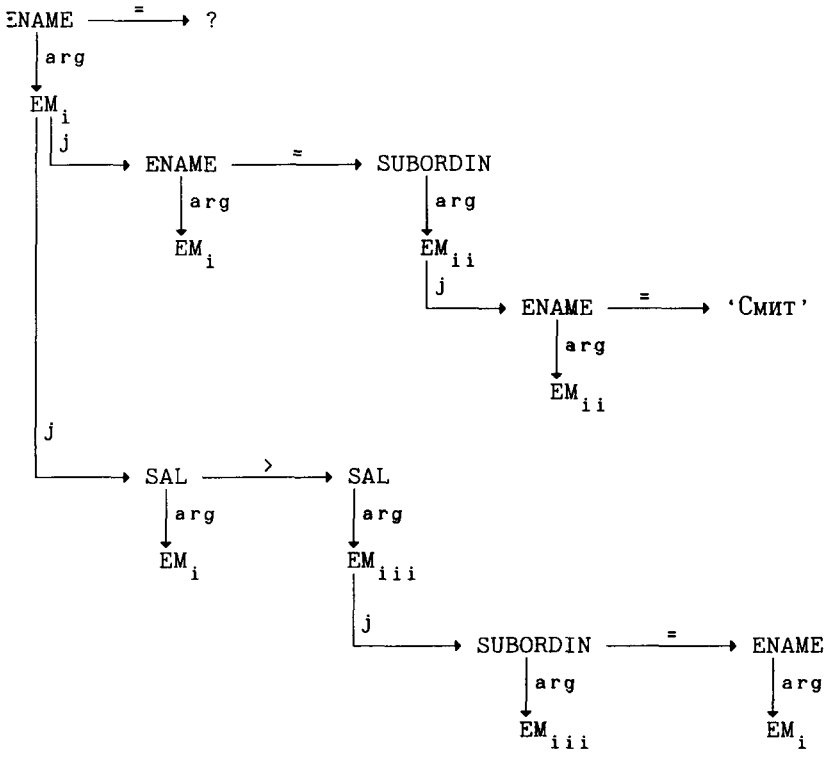
(3) Какие подчиненные Смита зарабатывают больше его начальника?




```
SELECT   ENAME
FROM     EM A
WHERE    (A.ENAME IN
                    (SELECT  SUBORDIN
                     FROM    EM B
                     WHERE   (B.ENAME = 'Смит')
                        AND (A.SAL >
                                    (SELECT  SAL
                                     FROM    EM D
                                     WHERE   (D.SUBORDIN =
                                       B.ENAME)))))
```

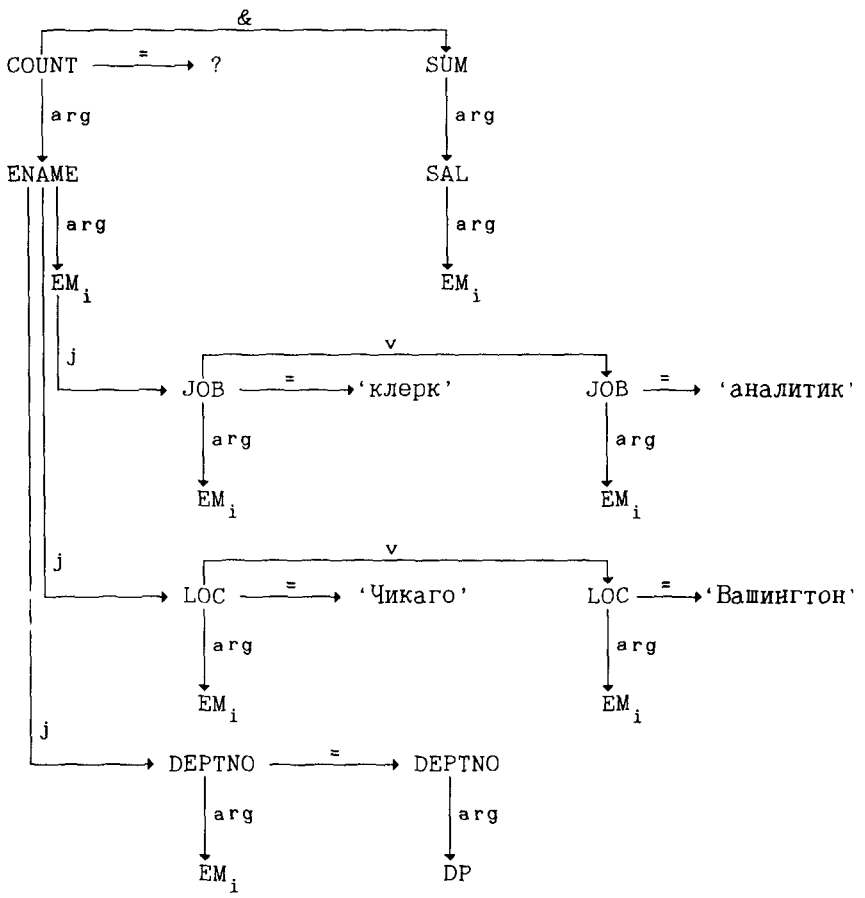
TIME:ØØ.2Ø

4) Какие подчиненные Смита зарабатывают больше своего начальника?




```
SELECT   ENAME
FROM     EM A
WHERE    (A.SAL > ALL
                 (SELECT  D SAL
                  FROM    EM D, EM A
                  WHERE   (D.SUBORDIN = A.ENAME)))
          AND (A.ENAME IN
                          (SELECT  SUBORDIN
                           FROM    EM E
                           WHERE   (E.ENAME = 'Смит')))
TIME:ØØ 25
```

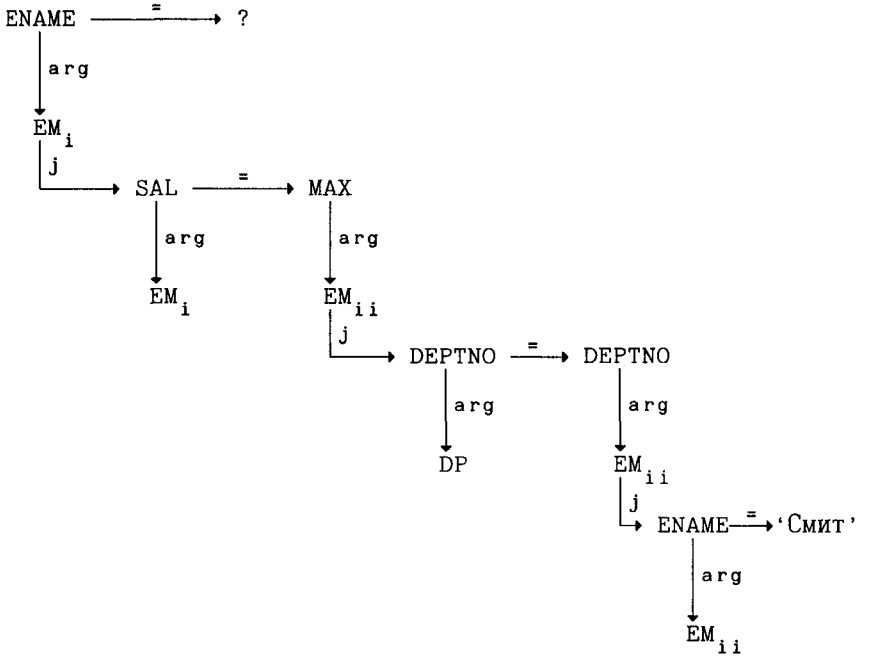
(5) Указать количество клерков и аналитиков, живущих в Чикаго и Вашингтоне, и их общую зарплату.




```
SELECT   COUNT(ENAME), SUM(SAL)
FROM     EM A, DP, CT
WHERE    (A.JOB = 'клерк' OR A.JOB = 'аналитик')
          AND (DP.LOC = 'Чикаго' OR CT.CITY = 'Вашингтон')
          AND (DP.DEPTNO = A.DEPTNO) AND (CT.CITY = DP.LOC)
```

TIME:ØØ.41

(6) Какие служащие получают зарплату, превышающую максимальную зарплату в отделе, в котором работает Смит?




```
SELECT   ENAME
FROM     EM A
WHERE    (A.SAL >
                 (SELECT   MAX(SAL)
                  FROM     EM, DP B
                  WHERE    (EM.DEPTNO = B.DEPTNO)
                    AND (B.DEPTNO IN
                                    (SELECT   DEPTNO
                                     FROM     EM C
                                     WHERE    (C.ENAME =
                                       'Смит')))))
```

TIME:ØØ.35

Каждый запрос сопровождается указанием времени (в минутах и секундах), которое потребовалось компьютеру для его обработки. Оценивая временные характеристики, следует иметь в виду, что в процессе работы компьютера к нему могли обращаться сразу несколько пользователей, что обычно увеличивало время обработки.

7.2. Эксперименты по русско-английскому МП

7.2.1. Вводные замечания

В результате интенсивной работы над лингвистическим процессором и над новой версией системы ЭТАП-2 авторам стало ясно, что требуется сравнительно немного усилий, чтобы превратить эту систему МП в двунаправленную и иметь возможность осуществлять перевод не только с английского языка на русский, но и с русского на английский.

В самом деле, алгоритмический и программный комплексы, используемые в системе, как мы неоднократно отмечали, универсальны и не зависят от того, какой конкретно естественный язык является входным или выходным. Самая сложная часть лингвистического обеспечения, требуемого для русско-английского перевода, - формальная модель синтаксиса русского языка (в частности, полный блок правил синтаксического анализа) - была готова и опробована в системе перевода с русского на SQL. По существу, из крупных блоков лингвистических правил нам недоставало правил нормализации русской синтаксической структуры, правил собственно перевода, правил развертывания английской синтаксической структуры и английского синтаксического синтеза.

Кроме того, было необходимо внести некоторые поправки в английский и русский комбинаторные словари системы ЭТАП-2. На этом участке возникло несколько неожиданных проблем, одну из которых мы обсудим ниже.

Начиная с 1978 г., когда авторы приступили к работе над своей первой (французско-русской) системой МП, предпринимались все усилия, чтобы полностью представить необходимую лингвистическую информацию в декларативном, а не в процедурном виде. Такая установка естественным образом приводит к тому, что входной и выходной языки получают независимые описания. Это, конечно, является принципиальным достоинством лингвистического обеспечения, потому что позволяет использовать одни и те же формальные модели языков в различных комбинациях для получения многих двуязычных систем МП, не говоря уже о том, что заданные таким образом лингвистические знания легко обновлять и корректировать.

Однако добиться полной независимости описания входного и выходного языков можно только при условии, что переход от языка к языку совершается на уровне универсальной СемС.

Поскольку в наших системах МП пока не удалось достичь такого уровня глубины и переход происходил на уровне нормализованной СинтС, оказалось желательным обеспечить какую-то степень соответствия входной и выходной СинтС для облегчения этапа собственно перевода. Это значит, что описания рабочих языков должны были в определенной мере соответствовать друг другу. Проще говоря, описание выходного языка в определенных точках следовало подогнать под описание входного.

Очевидно, что, давая определенный прагматический выигрыш в одном отношении, такая подгонка, в силу своей сугубо практической направленности, должна приводить и действительно приводит к некоторым нежелательным следствиям. Однако, начиная эту работу, мы не представляли себе их совокупного эффекта. Он стал ясен позднее, когда мы приступили к "обращению" лингвистического обеспечения англо-русской версии с целью построения русско-английской системы МП. Оказалось, что практически все решения, принимаемые для одного языка и для одного направления перевода, должны немедленно проверяться на "безопасность" по отношению к другому языку и к противоположному направлению перевода.

Сказанное можно проиллюстрировать следующим простым примером. В первоначальной англо-русской версии системы английское выражение *in the past* переводилось на русский язык выражением ***в прошлом***, которое для простоты трактовалось как безусловный оборот. Это решение было вполне приемлемо для однонаправленной системы перевода, поскольку русское выражение появлялось в самом конце работы системы и не могло повлиять на ее результат. Однако данное решение стало абсолютно недопустимым в двунаправленной системе МП: очевидно, что для целей синтаксического анализа трактовать ***в прошлом*** как единое слово нельзя. Это два слова, которые могут быть даже не связаны друг с другом непосредственно (как в выражении ***в прошлом году***), не говоря уже о том, что они могут разрываться другими словами (как в выражениях ***в далеком прошлом, в недавнем прошлом***).

В конечном счете нам пришлось пересмотреть данное и многие другие подобные решения. Как оказалось, ревизия таких решений привела к изменению моделей рабочих языков в сторону большей принципиальности. По нашему мнению, это обстоя-

тельство способствовало лучшему пониманию того, насколько важно избегать непринципиальных научных решений даже при работе над сугубо прикладными задачами.

Несмотря на все сказанное, процесс обращения англо-русской версии системы в русско-английскую все же потребовал от нас гораздо меньше времени и усилий, чем понадобилось бы в случае создания новой системы автоматического перевода.

Важно также подчеркнуть, что при этом обращении удалось сохранить абсолютно все логико-алгоритмическое и программное обеспечение системы. Поэтому при переходе от одного направления перевода к другому не требуется никакой дополнительной настройки; оператору достаточно лишь указать желаемый режим работы с помощью соответствующей команды.

После того, как все новые правила были написаны и необходимая модификация словарей произведена (включая насыщение русского КС словарными и трафаретными правилами нормализации и собственно перевода и насыщение английского КС словарными и трафаретными правилами развертывания синтаксической структуры и синтаксического синтеза), появилась возможность начать эксперименты с двунаправленной системой МП. Как было сказано выше, эта система получила название ЭТАП-3.

7.2.2. Основные параметры двунаправленной системы машинного перевода ЭТАП-3

Система ЭТАП-3 реализована на ЭВМ VAX-75Ø и работает практически на тех же научных и технологических принципах, что и ЭТАП-2.

Англо-русская версия системы ЭТАП-3, как и ее предшественница, имеет два основных режима - грубого пословного перевода, осуществляемого на уровне морфологической структуры предложения, и высококачественного пофразного перевода, который производится без вмешательства человека (т.е. не требует пред-, пост- и интерредактирования).

Русско-английская версия системы работает только в режиме высококачественного перевода. Внедрение режима грубого пословного перевода, которое могло бы преследовать чисто коммерческие цели, легко осуществимо и требует лишь модификации русского морфологического словаря.

Перевод предложения средней степени грамматической слож-

ности и средней длины (20-30 слов) занимает в обеих версиях системы от 20 до 80 сек (при работе ЭВМ в монорежиме).

В настоящее время морфологические словари системы ЭТАП-3 содержат 15 000 и 12 000 единиц (английский и русский словари соответственно), а комбинаторные словари - примерно по 10 000 единиц каждый.

Около 50 % словарных статей комбинаторных словарей представляют собой научно-техническую терминологию (из области электротехники и информатики), а остальная часть - общеупотребительную лексику. Словари обеспечивают удовлетворительное покрытие текста для экспериментальной эксплуатации, однако недостаточны для коммерческого использования системы.

Одной из сильных сторон системы ЭТАП-3 является высокая степень модульности. Как ясно из предшествующего изложения, лингвистическое обеспечение системы состоит из весьма значительного числа компонентов. В него входят два морфологических и два комбинаторных словаря (по одному для русского и английского языков), правила синтаксического анализа для русского и английского языков, правила собственно перевода, правила синтаксического синтеза для обоих языков и некоторые другие блоки правил. Тем не менее, четыре самых крупных компонента системы - морфологические и комбинаторные словари для обоих языков - удалось организовать таким образом, что они являются общими для обоих направлений перевода. При англо-русском переводе английские словари являются входными и служат для морфологического анализа, синтаксического анализа, нормализации и собственно перевода английского текста, а русские словари являются выходными и используются на этапах развертывания русской синтаксической структуры, синтаксического синтеза и морфологического синтеза русского текста. При русско-английском переводе входные и выходные словари меняются ролями[1].

---

[1]В порядке уточнения следует подчеркнуть, что если в морфологических словарях вся информация служит как для анализа, так и для синтеза (т.е используется в равной мере в обоих направлениях перевода), то в комбинаторных словарях картина несколько иная: хотя большая часть лингвистических сведений, содержащихся в классификационных зонах, является общей для обоих направлений перевода, некоторые другие классификационные и все операционные зоны словарной статьи (зона перевода, лексические правила анализа и собственно

Другие компоненты системы (синтагмы, блоки правил перевода и синтаксического синтеза и еще несколько вспомогательных блоков) используются только в одной версии, однако соответствующие компоненты для обоих направлений перевода построены на одних и тех же лингвистических принципах.

7.2.3. Новые свойства системы

По сравнению с предыдущей системой перевода - ЭТАП-2 - настоящая система обладает рядом важных преимуществ.

Одно из основных преимуществ состоит в активном использовании в процессе перевода описанного в разд. 4.8 синтаксического ускорителя - блока форсированного установления высоковероятных синтаксических связей. В настоящий момент этот блок используется только в русско-английской версии системы, однако в ближайшем будущем должен быть включен и в англо-русскую версию.

Напомним, что основная идея этого блока состоит в имитации "анализирующей" компетенции носителя языка, который обычно понимает (а, стало быть, и анализирует синтаксически) "бегущее" предложение почти в любой его точке. Аналогичным образом, синтаксический ускоритель, просматривая предложение слева направо, пытается установить максимально возможное число высоковероятных синтаксических связей в пределах относительно коротких фрагментов предложения.

Экспериментальная эксплуатация этого блока в синтаксическом компоненте системы ЭТАП-3 дает весьма обнадеживающие результаты. В среднем, как мы говорили, блок устанавливает около 80 % синтаксических связей, необходимых для формирования дерева синтаксической структуры предложения, а довольно часто и все 100 % связей, что приводит к существенному сокращению времени анализа и перевода в целом.

Следует оговориться, что синтаксический ускоритель работает с лингвистическими правилами, заданными в процедурном виде. На первый взгляд, этот факт является отступлением от наших принципов. Однако важно учесть, что при наличии готового декларативного описания лингвистической информации мы

перевода, правила нормализации, развертывания СинтС выходного языка и синтеза) предназначены только для одного направления перевода и активируются тогда, когда работает соответствующая версия системы.

можем существенно сократить время компьютерной лингвообработки, переписывая ее в процедурном виде.

Из других новшеств системы ЭТАП-3 стоит указать на принципиальную возможность обрабатывать сразу целый текст, не разбивая его на отдельные предложения вручную. Система сама, пользуясь информацией о естественных ограничителях предложения (знаках препинания), сегментирует текст на предложения и затем переводит их одно за другим.

Важным нововведением является также возможность осуществлять множественный перевод. Если до сих пор процесс перевода был детерминирован с самого начала, поскольку в ходе синтаксического анализа формировалась и подлежала дальнейшей переработке лишь первая из возможных синтаксических структур исходного предложения (даже в случае его синтаксической и/или лексической многозначности), то теперь мы можем вернуться ко всем "точкам ветвления" в процессе синтаксического анализа и получить несколько структур (и соответственно несколько переводов). Тем самым программно многовариантный МП обеспечен полностью.

Были проведены и первые лингвистические эксперименты на эту тему. Они вполне оправдали наши ожидания. В частности, для знаменитой фразы Н. Хомского Flying planes can be dangerous были получены три правильных перевода: 1) 'Управление самолетами может быть опасным', 2) 'Летающие самолеты могут быть опасными', 3) 'Летая, самолеты могут быть опасными'. Второй и третий переводы соответствуют одной и той же содержательной интерпретации, что естественно, поскольку реальных альтернативных осмыслений в данном случае ровно два.

7.2.4. Примеры работы русско-английской версии системы ЭТАП-3

В данном разделе будут кратко охарактеризованы результаты экспериментов по автоматическому переводу с русского языка на английский, проведенных авторами в конце 1990 - начале 1991 г.

В экспериментах по русско-английскому переводу мы работали с тремя основными типами текстов: 1) русским текстом статьи об электронном оборудовании, полученным системой ЭТАП-2 в процессе англо-русского перевода; таким образом, на этом тексте был, по существу, поставлен эксперимент по

обратному переводу (заинтересованный читатель может обратиться к уже упомянутой книге [Апресян и др., 1989] и сравнить текст обратного перевода с текстом оригинала); 2) текстами запросов к базе данных ORACLE; 3) чисто экспериментальными предложениями, предназначенными для лингвистической отладки анализа и перевода "трудных" морфологических характеристик, "трудных" синтаксических конструкций и "трудной" лексики.

Ниже приводятся некоторые образцы полученных переводов После каждого предложения, как и в разд. 7.1, указывается время перевода в минутах и секундах. Здесь также надо иметь в виду, что в процессе работы компьютера к нему могли обращаться сразу несколько пользователей, что часто замедляло время перевода.

7.2.4.1. Текст статьи об электронном оборудовании

(1) Будущее электроники в оборудовании → The future of electronics in equipment. TIME=ØØ 15

(2) Достижения, ожидаемые в спектроскопии, хроматографии, спектрометрии и электрохимии. → Advances expected in spectroscopy, chromatography, spectrometry and electrochemistry TIME=ØØ 2Ø

(3) В недавние годы правительственные постановления, технические требования, развитие микропроцессоров и множество других факторов вызвали переворот в химическом, биохимическом и промышленном оборудовании. → In recent years government decrees, technical demands, development of microprocessors and a set of other factors caused a revolution in chemical, biochemical and industrial equipment TIME=Ø1 2Ø

(4) Тридцать пять лет назад значительные успехи были достигнуты в оборудовании химического анализа, и спектрометры, масс-спектрометры и водородные счетчики становятся доступными для массового пользователя. → Thirty five years ago significant successes were achieved in equipment of chemical analysis, and spectrometers, mass-spectrometers and hydrogen counters become accessible to mass user TIME=Ø1 16

(5) Поиск синтетического заменителя натурального каучука, например, ускорился в течение второй мировой войны и

стимулировал развитие оборудования, которое могло анализировать молекулярную структуру в лаборатории. → Search for synthetic substitute of natural rubber, for example, has accelerated during the second world war and stimulated the development of the equipment which could analyse a molecular structure in laboratory TIME=ØØ 5Ø

(6) В начале шестидесятых годов хроматография стала основным аналитическим инструментом, в настоящее время наиболее широко используемым в химических лабораториях. → At the beginning of the sixties chromatography became a basic analytical instrument at present most widely used in chemical laboratories TIME=ØØ 49

(7) Создание усовершенствованных детекторов и техники выборки требует более экзотических инструментальных методов, таких как масс-спектрометрия, газовая хроматография, рентгеновская флуоресценция и плазменная спектроскопия. → Creation of improved detectors and a technique of sampling requires more exotic instrumental methods such as mass-spectrometry, gas chromatography, X-ray fluorescence and plasma spectroscopy TIME=ØØ 36

(8) Стыковка этих методов приводит к еще более мощным инструментам. → Interfacing these methods results in even more powerful instruments TIME=ØØ 19

(9) Исследования в биологии и медицине в двадцатом веке произвели переворот в приборостроении, в значительной мере определивший успех этой области. → Investigations in biology and medicine in the twentieth century have produced a revolution in instrument-making industry largely defining a success of this region TIME=ØØ 37

(10) Полимерная химия, охватывающая производство всех пластмасс, не могла бы продвинуться к своему современному состоянию без инструментов, разработанных в течение последних пятидесяти лет. → Polymeric chemistry embracing the production of all plastics couldn't advance to its modern state without instruments developed during the last fifty years TIME=ØØ 44

(11) Несколько лет назад стали коммерчески доступны первые микропроцессорные схемы, что привело к созданию многочисленных вариантов машины ответа. → Several years ago became commercially accessible the first microprocessor

chips, which has led to creation of numerous versions of a machine of answer TIME=Ø1 14

(12) Эти приборы обеспечили исследователям автоматическое считывание названий компонентов и концентраций, не требуя инверсии сложных матриц. → These devices have provided researchers with automatic reading of the names of components and concentrations not requiring an inversion of complex matrices TIME=ØØ 27

(13) В прошлом спектрографический анализ требовал точного измерения многокомпонентных образцов хорошо обученными специалистами. → In the past the spectrographic analysis required a precise measurement of multicomponent samples by well trained specialists TIME=ØØ 31

(14) Из-за неизбежного перекрытия спектров поглощения оператор обычно должен был решать математические уравнения для определения относительных концентраций многокомпонентной смеси. → Because of unavoidable overlapping of absorption spectra the operator usually had to solve mathematical equations for definition of relative concentrations of multicomponent mixture TIME=Ø1 ØØ

(15) Если образец содержал более десяти компонентов, обычно требовалась подготовка и разделение сложных образцов, чтобы устранить помехи или чтобы обнаружить следы. → If the sample contained more than ten components usually preparation and a separation of complex samples were required to eliminate interferences or to detect traces TIME=ØØ 43

(16) Микропроцессоры устранили необходимость изучения почти незаметных пиков поглощения для подтверждения наличия специальных соединений. → Microprocessors have eliminated the necessity of learning almost indistinguishable peaks of absorption for confirmation of the presence of special connections TIME=ØØ 25

(17) Это также устранило необходимость в утомительных исправлениях основной линии или математических преобразованиях матриц для определения концентраций. → It has also eliminated the necessity of tedious corrections of the basic line or mathematic conversions of matrices for definition of concentrations TIME=ØØ 36

(18) Улучшение автоматизации спектрофотометров посред-

ством автоматических систем выборки и обработки данных бросает вызов спектроскопии в восьмидесятые годы. → The improvement of automation of spectrophotometers by automatic systems of sampling and data processing is a challenge to spectroscopy in the eighties TIME=ØØ 35

(19) Возможности качественного анализа спектроскопии будут комбинироваться с газовой хроматографией и высокоэффективной жидкостной хроматографией для обеспечения специфичности, необходимой, чтобы точно измерять компоненты следа → Possibilities of a qualitative analysis of spectroscopy will be combined with gas chromatography and highly effective liquid chromatography for support of specificity, necessary to accurately measure components of trace TIME=Ø1 23

(20) Приборы, покрывающие целый диапазон ультрафиолетовых лучей посредством волн инфракрасного спектра, будут управляться микропроцессорной техникой → Devices, covering the entire range of ultraviolet rays by waves of infrared spectrum, will be controlled by microprocessor engineering TIME=ØØ 32

(21) В плазменной эмиссионной спектроскопии, которая является относительно новым методом спектрального анализа, эмиссионные линии компонентов образца контролируются, когда образец сжигается в плазме → In plasma emission spectroscopy, which is a relatively new method of a spectral analysis, emission lines of sample components are monitored, when the sample is burned in plasma TIME=Ø1 ØØ

(22) Методы плазменной спектроскопии включают прямоточную плазму, индуктивно соединенную плазму и индуцированную микроволновую плазму → Methods of plasma spectroscopy comprise direct current plasma, inductively bound plasma and induced microwave plasma TIME=ØØ 28

(23) Многочисленные фотоумножительные детекторы устанавливаются за отверстиями, помещенными в оптически рассеянном спектре, и одновременно производится анализ сразу 30 элементов и возможных соединений → Numerous photomultiplier detectors are set behind holes, placed in optically dispersed spectrum, and simultaneously analysis is produced of at once 3Ø elements and possible connections TIME=ØØ 48

(24) В таких приложениях, как анализ качества воды и

обнаружение металлов в смазочных маслах, эмиссионная спектроскопия быстро заменяет методы атомного поглощения. → In such applications as an·analysis of water quality and detection of metals in lubricating oils emission spectroscopy quickly replaces methods of atomic absorption TIME=ØØ 55

(25) Это происходит, поскольку в эмиссионной спектроскопии не возникает необходимости менять источники для каждого измерения элемента. → It happens, since in emission spectroscopy doesn't arise the necessity to change sources for each measurement of an element TIME=ØØ 37

(26) В восьмидесятые годы появляется возможность комбинировать эмиссионную спектроскопию и атомное поглощение в одном блоке, снабженном автоматической выборкой и дисплеем для более широкого диапазона анализа. → In the eighties appears a possibility to combine emission spectroscopy and atomic absorption in single block, supplied with automatic sampling and display for wider range of analysis TIME=ØØ 59

(27) В настоящее время микропроцессоры упростили спектрографический анализ, и многокомпонентные образцы могут анализироваться обычным образом, не требуя последующих расчетов или интерпретации оператором. → At present microprocessors have simplified a spectrographic analysis, and multicomponent samples can be analyzed routinely, not requiring subsequent calculations or interpretation by the operator TIME=ØØ 58

(28) Ядерное оборудование, которое появилось в конце шестидесятых годов, изменило медицинскую диагностику. → The nuclear equipment, which appeared in the end of the sixties, has changed medical diagnostics TIME=ØØ 24

(29) Совершенствование спектрофотометров посредством автоматических систем выборки и обработки данных бросает вызов спектроскопии в восьмидесятые годы. → Improvement of spectrophotometers by automatic systems of sampling and data processing is a challenge to spectroscopy in the eighties TIME=ØØ 29

7.2.4.2 Запросы к демонстрационной базе данных ORACLE

(30) Привести список имен, должностей и зарплат служащих, имеющих должность и зарплату, как у Елены Форд. → Give

a list of names, jobs and salaries of employees having a job and salary, as Helen Ford has TIME=ØØ 34

(31) Указать фамилию, зарплату и место работы каждого менеджера. → Specify the last name, salary and office of each manager TIME=ØØ 12

(32) Кто в отделе сбыта получает максимальную зарплату? → Who in the commercial department gets the maximum salary? TIME=ØØ 17

(33) Каков размер максимальной зарплаты в отделе сбыта? → What is the size of the maximum salary in the commercial department? TIME=ØØ 18

(34) Кто из менеджеров бухгалтерии получает зарплату выше средней зарплаты клерков отдела сбыта? → Who of the managers of the accounting department gets a salary above the average salary of clerks of the commercial department? TIME=ØØ 27

(35) Кто из менеджеров отдела сбыта получает максимальную зарплату? → Who of the managers of the commercial department gets the maximum salary? TIME=ØØ 12

(36) Как зовут служащего, который работает клерком в отделе сбыта? → What is the name of the employee, who works as a clerk in the commercial department? TIME=ØØ 18

(37) Какую зарплату платят Джоунзу из отдела сбыта? → What salary is paid to Jones from the commercial department? TIME=ØØ 18

(38) Привести среднюю зарплату по каждому отделу, в котором имеется по два аналитика. → Specify the average salary over each department, in which there are two analysts TIME=ØØ 27

(39) Привести список всех отделов, имеющих по меньшей мере двух клерков. → Give a list of all departments having at least two clerks TIME=ØØ 23

(40) Указать клерков, зарплата у которых выше, чем у Джона. → Specify the clerks, whose salary is higher than that of John TIME=ØØ 21

(41) Средняя зарплата клерков каких отделов составляет 300 рублей? → Is the average salary of clerks of what departments 3ØØ roubles? TIME=ØØ 21

(42) Сколько клерков работает в коммерческом отделе, и какова их суммарная зарплата? → How many clerks work in the

commercial department, and what is their total salary" TIME=ØØ.2Ø.

(43) Для каждой должности подсчитать количество служащих, занимающих эту должность. → For each job count the number of employees occupying this job. TIME=ØØ.24.

(44) Найдите всех служащих, имеющих должность менеджера или работающих клерками в отделе сбыта. → Find all employees having a job of a manager or working as clerks in the commercial department. TIME=ØØ.25.

(45) Указать служащих, зарабатывающих больше кого-нибудь из аналитиков. → Specify the employees earning more than somebody of the analysts. TIME=ØØ.22.

(46) Найдите всех служащих в бухгалтерии, имеющих такие же зарплаты, как у служащих отдела сбыта. → Find all employees in the accounting department having the same salaries as employees of the commercial department have. TIME=ØØ.54.

(47) Какой отдел возглавляется служащим, зарабатывающим три тысячи долларов? → What department is headed by the employee earning three thousand dollars? TIME=ØØ.16.

(48) Какие города находятся в штате, который называется Калифорния? → What cities are in the state, which is called California? TIME=ØØ.37.

(49) Перечислите отделы, которые находятся там же, где бухгалтерия. → List the departments, which are in the same place as the accounting department. TIME=ØØ.35.

(50) Перечислите отделы, которые находятся там же, где находится бухгалтерия. → List the departments, which are in the same place as is the accounting department. TIME=ØØ.25.

(51) Какие отделы имеются в штате, в котором расположен Даллас? → What departments are there in the state, in which Dallas is situated? TIME=ØØ.28.

(52) Кто и на какой должности работает в коммерческом отделе? → Who and on what job work in the commercial department? TIME=ØØ.19.

7.2.4.3. Экспериментальные предложения

(53) Проблема не будет рассматриваться сегодня. → The problem won't be considered today. TIME=ØØ.13.

(54) Ни одна машина не работала. → Not a single machine worked. TIME=ØØ.12.

(55) Программа занимает слишком много места в памяти

компьютера. → The program occupies too much place in memory of computer. TIME=ØØ.17.

(56) Я сразу же увидел две толстые книги, лежащие на столе. → I at once saw two thick books lying on table. TIME=ØØ.32.

(57) Где расположен город Нью-Йорк? → Where is the city of New York situated? TIME=ØØ.12.

(58) Кем работают Иван и Петр? → What is the job of Ivan and Peter? TIME=ØØ.12.

(59) Иван занял у меня немного денег. → Ivan has borrowed from me a little money. TIME=ØØ.13.

(60) Он стоял слева от меня. → He stood on my left. TIME=ØØ.11.

(61) Иван должен решать свои задачи. → Ivan must solve his tasks. TIME=ØØ.10.

(62) Иван должен был решать задачи. → Ivan had to solve tasks. TIME=ØØ.18.

(63) Она не должна была решать эту задачу. → She hadn't to solve this task. TIME=ØØ.20.

(64) Она не должна решать эту задачу. → She must not solve this task. TIME=ØØ.17.

(65) Она не должна будет решать эту задачу. → She won't have to solve this task. TIME=ØØ.21.

(66) Решить задачи хотел Петр. → It was Peter that wanted to solve tasks. TIME=ØØ.10.

(67) Решить эту задачу мог только Игорь. → It was only Igor that could solve this task. TIME=ØØ.11.

(68) Решать такие задачи должно было именно новое устройство. → It was just the new apparatus that had to solve such tasks. TIME=ØØ.15.

(69) Игорь пошел бы в институт, если бы у него было время. → Igor would go into institute if he had time. TIME=ØØ.26.

(70) Джон отсутствовал на работе вчера. → John was absent at work yesterday. TIME=ØØ.15.

(71) Джон, отсутствовавший в Москве в течение двух лет, приехал сегодня. → John, absent in Moscow during two years, came today. TIME= ØØ.25.

(72) Сколько воды выпил Джон? → How much water has John drunk? TIME=ØØ.08.

(73) Два человека уже пришли. → Two men have come already. TIME= ØØ.Ø9.

(74) Джон еще не пришел. → John hasn't come yet TIME=ØØ.Ø9.

(75) Как зовут вашу жену? → What is the name of your wife? TIME= ØØ.1Ø.

(76) Сколько лет вашей жене? → How old is your wife? TIME=ØØ.14.

(77) Сколько лет было вашей жене, когда она работала в Чикаго? → How old was your wife, when she worked in Chicago? TIME=ØØ.21.

(78) Джону был один год. → John was one year old. TIME=ØØ.11.

(79) Смит знал, когда пришел Джоунз. → Smith knew, when Jones came. TIME=ØØ.17.

(80) И Иван, и Петр получают большую зарплату. → Both Ivan and Peter get a big salary. TIME=ØØ.15.

(81) Иван работал с 1980 по 1990 год. → Ivan worked from 198Ø to 199Ø. TIME=ØØ.18.

(82) Мы работаем с понедельника по пятницу. → We work from Monday to Friday. TIME=ØØ.12.

(83) Сколько денег платят Игорю за его работу? → How much money is being paid to Igor for his work? TIME=ØØ.13.

ЗАКЛЮЧЕНИЕ

В предстоящие пять лет ЛП должен пройти путь от экспериментальной до реально действующей системы.

В рамках подсистем общения с базами данных на неограниченном естественном языке нам предстоит развить ЛП в следующих направлениях: 1) расширить экспериментальную базу ЛП за счет включения новых типов данных, с перспективой перехода от демонстрационной базы к реальной; 2) расширить круг запросов на SQL, доступных для ЛП; 3) расширить круг средств ЕЯ, доступных для ЛП; 4) провести оптимизацию инструментальных средств ЛП для ускорения реакции системы на запрос; 5) обогатить систему средствами контроля правильности понимания запросов, включив в нее, как было предусмотрено в главе 5, подсистему синтеза проанализированного запроса: на вход синтеза должна поступать та СемС, которая получилась в результате анализа, а на выходе должна получаться более эксплицитная по сравнению с первоначальной формулировка запроса на ЕЯ.

В рамках систем МП предстоит прежде всего существенно расширить словари и, возможно, внести некоторые дополнения в массивы правил.

Кроме того, системы МП тоже должны быть оснащены механизмами, дающими более твердые, чем в нынешней системе, гарантии того, что система правильно поняла и перевела текст. В сущности, таким механизмом является возможность многовариантного перевода омонимичных предложений: если первый выданный перевод не удовлетворяет пользователя, по его просьбе будет выдан второй, третий и т. д., вплоть до исчерпания всех альтернатив. Как мы говорили, программно такая возможность у нас уже есть, т. е. мы реально получаем для переводимого предложения не одну синтаксическую структуру, как в предшествующей версии ЭТАПа, а все омонимичные структуры (см. в гл. 7 переводы для фразы Н. Хомского Flying planes can be dangerous). Очевидно, однако, что есть такие формально омонимичные предложения, которым не отвечают какие-либо разумные альтернативные интерпретации, т. е.

такие, каждая из которых соответствует какому-то фактически мыслимому положению вещей. В связи с этим возникает лингвистическая задача изучения тех типов синтаксической омонимии, которые реально допускают возможность различных интерпретаций. Именно для таких типов в системе должны быть предусмотрены все лингвистические правила, обеспечивающие многовариантность перевода соответствующих предложений.

Еще одно новое направление исследований связано с перспективой интеграции англо-русской версии ЭТАПа-3 в качестве вспомогательной подсистемы в систему перевода текстов запросов с ЕЯ на SQL, в результате чего эта последняя станет двуязычной. При этом русский язык, который уже используется в качестве естественного средства общения с базами данных, должен послужить языком-посредником между английским и SQL. Уже проведенные эксперименты свидетельствуют о том, что в этой подсистеме можно будет переходить от английского языка к русскому на уровне нормализованной структуры, т.е. далеко не достигая уровня реального русского предложения. При этом время обработки английского запроса оказывается почти таким же, как для аналогичного русского.

В ближайшие планы авторов системы входит и усовершенствование процедуры обработки предложений, содержащих произвольное число неопознанных слов в произвольной синтаксической позиции. В настоящее время блок синтаксического анализа удовлетворительно обрабатывает лишь небольшой процент предложений с неопознанными словами, позиция которых благоприятствует правильной идентификации их синтаксической функции, а в остальных случаях отказывается строить синтаксическую структуру и переходит к дежурному режиму пословного перевода. Нам предстоит разработать механизмы идентификации синтаксической функции неопознанных элементов в гораздо более широком круге ситуаций, чтобы получать идиоматичный перевод даже в тех случаях, когда эти элементы стоят в синтаксически неблагоприятных для идентификации позициях.

Проведенная с ЛП работа убедила нас в том, что различные его лингвистические компоненты могут использоваться и за пределами тех автоматических информационных систем, для которых он первоначально предназначался. Одно из возможных приложений его словарных компонентов - партнерские компьютерные системы типа "помощник учителя русского и иностран-

ных языков". Они мыслятся как системы овладения языком или совершенствования языковых навыков человека в активном игровом режиме. Концепция такого "помощника" уже разработана. Есть основания надеяться, что на этом пути удастся создать эффективное компьютерное средство, помогающее практически овладеть морфологией языка, весьма значительной частью лексики, включая многочисленные синонимы, антонимы, конверсивы и дериваты данного слова и правила его сочетаемости с другими словами, а также некоторыми элементами синтаксиса (преимущественно моделями управления).

Таким образом, если попытаться кратко сформулировать перспективы дальнейшей работы над ЛП, то можно будет сказать, что нашей целью является создание таких теоретических моделей естественных языков, которые являются компонентами многоязычного и полифункционального ЛП.

Литература

Альшванг В.Д., Коровина Т.И., Кузьменчук В.А., Сергеев А.С., Сиротюк Г.И., Цинман Л.Л. Математическое обеспечение системы автоматического перевода ЭТАП-1 // Прикладные и экспериментальные лингвистические процессоры Новосибирск ВЦ СО АН СССР, 1982 С 3-25

Апресян Ю.Д. Синтаксические признаки лексем // Russ. Linguist. 1985. Vol. 9, N 2/3. P. 289-318.

Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л., Лазурский А.В., Перцов Н.В., Санников В.З., Цинман Л.Л. Лингвистическое обеспечение системы ЭТАП-2 М Наука, 1989 295 с

Апресян Ю.Д., Цинман Л.Л. Об идеологии системы ЭТАП-2 // Формальное представление лингвистической информации Новосибирск ВЦ СО АН СССР, 1982 С 3-19

Богуславский И.М. Лингвистический процессор и локативные обстоятельства //Вопр языкознания 1991 № 1 С 69-78

Богуславский И.М., Цинман Л.Л. Семантический компонент лингвистического процессора // Семиотика и информатика. 1990а Вып 30 С 5-30

Богуславский И.М., Цинман Л.Л. На естественном языке? // Мы и компьютер 1990б № 9 С 11-12

Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения М Наука, 1985 143 с

Грамматика русского языка М Изд-во АН СССР, 1960 Т 1 Фонетика и морфология 719 с Т 2 Синтаксис Ч 1 702 с Ч 2 440 с

Грамматика современного русского литературного языка М Наука, 1970 767 с

Диненберг Е.Г., Кучин С.И., Трапезников С.П. INTERBASE - система конструирования ЕЯ-интерфейса к dBASE III PLUS // Всесоюз конф "Искусственный интеллект - 90" Минск, 1990 С 161-164

Еськова Н.А., Мельчук И.А., Санников В.З. Формальная модель русской морфологии М , 1971 71 с (Препр /ИРЯ АН СССР)

Железняков М.М., Крупко Н.А. Реализация синтаксической модели зависимостей с учетом упорядочения связей и предпочтений // Формальное описание структуры естественного языка Новосибирск ВЦ СО АН СССР, 1980 С 94-104

Зализняк А.А. Русское именное словоизменение М Наука, 1967 370 с

Иомдин Л.Л. Автоматическая обработка текстов на естественном языке Модель согласования М Наука, 1990 168 с

Иомдин Л.Л., Мельчук И.А., Перцов Н.В. Фрагмент модели русского поверхностного синтаксиса 1 Предикативные синтагмы // НТИ Сер 2, Информ процессы и системы 1975 № 7 С 30-43

Иомдин Л.Л., Перцов Н.В. Фрагмент модели русского поверхностного синтаксиса 2 Комплетивные и присвязочные конструкции // Там же 1975 № 11 С 22-32

Коровина Т.И., Румшиский Б.Л., Семенова В.Э., Цинман Л.Л. Аппарат для описания морфологии флективных языков М , 1977 45 с (Препр / ИРЯ АН СССР, № 91)

Крупко Н.А., Цейтин Г.С. Разработка языкового процессора для системы управления // Взаимодействие с ЭВМ на естественном языке Новосибирск ВЦ СО АН СССР, 1978 С 147–156

Кулагина О.С. Исследования по машинному переводу М Наука, 1979 320 с

Кулагина О.С. Морфологический анализ русских глаголов М , 1985 28 с (Препр / ИПМ АН СССР, № 195)

Кулагина О.С. Об автоматическом синтаксическом анализе русских текстов М , 1987 22 с (Препр / ИПМ АН СССР, № 205)

Кулагина О.С. Машинный перевод современное состояние // Семиотика и информатика 1989 Вып 29 С 5–34

Кулагина О.С. О синтаксическом анализе на основе предпочтений М , 1990 20 с (Препр / ИПМ АН СССР, № 3)

Лазурский А.В., Митюшин Л.Г., Санников В.З., Цинман Л.Л. Морфологический компонент лингвистического процессора // Всесоюз конф по искусств интеллекту Тез докл Переславль-Залесский, 1988 Т 3 С 119–124

Лейкина Б.М., Цейтин Г.С. Синтаксическая модель с допущением ограниченной непроективности // Международный семинар по машинному переводу М ВЦП, 1975 С 72–74

Мельчук И.А. Автоматический синтаксический анализ 1 Внутрисегментный анализ Новосибирск Наука, 1964 357 с

Мельчук И.А. Опыт теории лингвистических моделей "Смысл ↔ Текст" М Наука, 1974 314 с

Митюшин Л.Г. Длина синтаксических связей и индуктивные структуры // Семиотика и информатика 1985 Вып 26 С 34–51

Митюшин Л.Г. О высоковероятных синтаксических связях // Проблемы разработки формальной модели языка Сер "Вопросы кибернетики" Вып 137 М Научный совет по комплексной проблеме "Кибернетика" АН СССР, 1988 С 145–174

Пешковский А.М. Русский синтаксис в научном освещении М Учпедгиз, 1956 511 с

Райхлина А.М., Цинман Л.Л. Алгоритмическая процедура "Работа с правилом преобразования" // Предварительные публикации М ИРЯ АН СССР 1986 Вып 174 С 29–46

Рахилина Е.В. Семантика локативных вопросов (вопросы со словом ГДЕ) // Проблемы разработки формальной модели языка Сер "Вопросы кибернетики" Вып 137 М Научный совет по комплексной проблеме "Кибернетика" АН СССР, 1988 С 87–99

Русская грамматика М Наука, 1980 Т 1 Фонетика Фонология Ударение Интонация Словообразование Морфология 783 с Т 2 Синтаксис 709 с

Саввина Е.Н. Фрагмент модели поверхностного синтаксиса русского языка Сравнительные конструкции (сравнительные и отсоюзные синтагмы) // НТИ Сер 2, Информ процессы и системы 1976 № 1 С 38–43

Санников В.З. Правила сочетаемости членов русских сочинительных конструкций (СК) // Linguist. silesiana, 1981. Vol. 4. P 111-131.

Санников В.З. Русские сочинительные конструкции Семантика Прагматика Синтаксис М Наука, 1989 267 с

Симонс Дж. ЭВМ пятого поколения компьютеры 90-х годов М Финансы и статистика 1985 172 с

Ульман Дж. Основы систем баз данных М Финансы и статистика, 1983 334 с

Урысон Е.В. Поверхностно-синтаксическое представление русских аппозитивных конструкций // Wien. Slaw. Almanach 1981. Bd. 7. S. 155-215.

Урысон Е.В. Направление синтаксической зависимости в русских аппозитивных конструкциях // Bull. Soc. pol. linguist. 1982. Fasc. 39. P. 91-107.

Цейтин Г.С. Методы синтаксического анализа, использующие предпочтение языковых конструкций: модели и эксперименты // Международный семинар по машинному переводу. М.: ВЦП. 1975. С. 131-133.

Цинман Л.Л. Язык для записи лингвистической информации в системе автоматического перевода ЭТАП (опыт "практической логики") // Семиотика и информатика. 1986а. Вып. 27 С. 82-120.

Цинман Л.Л. Развитие логико-алгоритмического обеспечения во второй очереди системы ЭТАП // Предварительные публикации. М.: ИРЯ АН СССР, 1986б. Вып. 174. С. 5-17.

Цинман Л.Л., Райхлина А.М., Розанова В.А., Сиротюк Г.И., Митюшин Л.Г. Алгоритм синтаксического анализа в системе ЭТАП-2 // 1986. Там же. С. 18-28.

Эршил У. СУБД завтрашнего дня, доступные сегодня // В мире персон. компьютеров. 1988, № 1.

Boguslavskij I.M., Tsinman L.L. A linguistic processor for natural language dialogue with databases // Artificial Intelligence and Information-Control System of Robots - 89. / Ed. I. Plander. Amsterdam: Elsevier Science Publishers B.V. 1989. p. 273-276.

Boguslavskij I.M., Tsinman L.L. Semantics in a linguistic processor // Computers and Artif. Intelligence. 1991. N. 3. P. 3-20.

Carcagno D., Iordanskaja L. Text planning in a generation report system // Odyssee recherches appliquees 1290. Montreal: Van Horne, 1988.

Chomsky N. Three models for the description of language // IRE Trans. Inform. Theory. 1956, Vol. 2, N. 3. P. 113-124.

Chomsky N. Syntactic structures. s'Gravenhage, 1957. 116 p.

Computational Linguistics 1985. Vol. 11, N. 1-3.

Ginsparg J.M. A robust portable natural language data base interface // Proc. of the 1983 conf. on appl. natural language processing. 1983.

Hovy E.F. On the study of text planning and realization // AAAI workshop on text planning. St. Paul, 1988.

Iordanskaja L.N., Polguere A. Semantic processing for text generation // Proc. of the Intern. comput. sci. conf. Hong Kong, 1988.

Kahn S. An overview of three relational data base products // IBM Syst. J. 1984. Vol. 23, N 2.

Lehmann H., Ott N., Zoeppritz M. Multilingual interface to databases // IEEE Database Eng. Bull. 1985. N 8.

Machine translation: Theoretical and methodological issues / Ed. S. Nirenburg. Cambridge: Cambridge Univ. Press, 1987. 350 p.

Maruyama N., Morohashi M., Umeda S., Sumita E. A Japanese sentence analyzer // IBM J. Res. and Develop. 1988. Vol. 32, N 2.

Maruyama N., Watanabe A. A discourse analysis technique for a natural language interface system // Proc. of the COMPSAC (The eleventh annual Intern. comput. software

and appl. conference) / IEEE Comput. Soc. Tokyo, 1987.
Mel'čuk I.A., Pertsov N.V. Surface syntax of English: A formal model within the Meaning ↔ Text framework . Amsterdam (Philadelphia): Benjamins, 1987. 527 p. (Linguist. and Lit. Stud. East. Europe, Vol. 13).
Sordi J.J. The query management facility // IBM Syst. J. 1984. Vol. 23, N 2. P. 126-150.
Tesnière L. Eléments de syntaxe structurale. P. 1959. 670 p.
Tsujii J., Muto Y., Ikeda Y., Nagao M. How to get preferred readings in natural language analysis // Proceedings of COLING 88. Budapest, 1988. Vol. 2. P. 683-687.
Waterman D.A. A guide to expert systems. Addison-Wesley, 1986. 419 p.

ОГЛАВЛЕНИЕ

Глава 1. Общее представление о лингвистическом процессоре: структура, назначение и принципы разработки . . 3
1.1. Постановка задачи . . . 3
1.2. Структура и состав ЛП . . . 5
1.3. Возможные прикладные функции ЛП . . . 8
1.4. Принципы разработки ЛП . . . 11
1.5. Источники разработки и компоненты ЛП . . . 13

Глава 2. Формальные языки для записи лингвистической информации . . . 16
2.1. Структура правила . . . 17
2.2. Сигнатура формального языка . . . 18
2.2.1. Термы . . . 18
2.2.1.1. Предметные константы . . . 18
2.2.1.2. Предметные переменные . . . 19
2.2.2. Предикаты . . . 20
2.2.2.1. Элементарные предикаты . . . 20
2.2.2.2. Составные предикаты . . . 24
2.2.3. Инструкции и параметры . . . 25
2.2.4. Анонимность неповторимых переменных . . . 25
2.3. Запись условий в зоне проверки . . . 25
2.3.1. Дизъюнктивная нормальная форма (ДНФ) . . . 25
2.3.2. Необходимые и невозможные условия . . . 26
2.3.3. Группы условий . . . 26
2.3.4. Выделенная переменная X . . . 26
2.3.5. Выделенная переменная Y . . . 26
2.4. Запись инструкций в зоне действий . . . 27
2.5. Дополнительные указания алгоритму, задаваемые в правилах . . . 30
2.5.1. Указатель возможной непроективности . . . 30
2.5.2. Указатель, управляющий обходом структуры . . . 30
2.6. Параметры трафаретных правил . . . 31

Глава 3. Формальная модель морфологии . . . 32
3.1. Понятие морфологической структуры . . . 32
3.2. Общие сведения о модели морфологии . . . 34
3.3. Морфологические признаки и характеристики . . . 37
3.3.1. Вводные замечания . . . 37
3.3.2. Список русских морфологических признаков . . . 38
3.4. Запись информации в морфологическом словаре . . . 40
3.4.1. Лексемы и безусловные обороты . . . 41
3.4.2. Стандартные объекты . . . 44
3.4.3. Образцы словарных статей морфологического словаря . . . 47
3.5. Стандартные объекты в описании русской морфологии . . . 48
3.5.1. Стандартные списки . . . 49
3.5.2. Стандартные ограничители . . . 51
3.5.3. Примеры форматов и трафаретов . . . 52
3.6. Алгоритмы морфологического анализа и синтеза . . . 53

Глава 4. Формальная модель синтаксиса . . . 56
4.1. Постановка задачи . . . 56
4.2. Понятие синтаксической структуры предложения . . . 57
4.3. Синтаксические отношения (комментированный перечень) . . . 60
4.3 1. Актантные СинтО . . . 61
4.3.2. Атрибутивные СинтО . . . 65
4.3.3. Сочинительные СинтО . . . 70
4.3.4. Служебные СинтО . . . 71

4.4. Принципиальная схема синтаксического анализа..... 73
4.5. Предсинтаксический анализ..... 74
4.5.1. Комбинированная морфологическая структура предложения..... 74
4.5.2. Назначение блока предсинтаксического анализа... 76
4.5.3. Образцы правил предсинтаксического анализа..... 78
4.5.3.1. Общее правило предсинтаксического анализа.. 78
4.5.3.2. Трафаретное правило предсинтаксического анализа..... 79
4.5.3.3. Словарное правило предсинтаксического анализа..... 80
4.6. Синтаксический анализ в собственном смысле..... 80
4.6.1. Типы синтаксических правил. Понятие синтагмы... 81
4.6.2. Образцы синтагм..... 82
4.6.2.1. Общая синтагма..... 82
4.6.2.2. Трафаретная синтагма..... 84
4.6.2.3. Словарная синтагма..... 86
4.7. Правила синтаксического синтеза..... 88
4.7.1. Общее правило..... 88
4.7.2. Трафаретное правило..... 90
4.7.3. Словарное правило..... 92
4.8. Алгоритмы синтаксического анализа..... 92
4.8.1. Основной алгоритм..... 93
4.8.1.1. Алгоритмическая процедура работы с правилом преобразования..... 93
4.8.1.2. Алгоритм предсинтаксического анализа..... 94
4.8.1.3. Алгоритм синтаксического анализа..... 94
4.8.2. Алгоритм установления высоковероятных связей.. 102
4.8.2.1. Общие сведения..... 102
4.8.2.2. Алгоритм установления связей..... 104
4.8.2.3. Синтагмы..... 107

Глава 5. Формальная модель семантики..... 113
5.1. Возможные пути построения модели..... 113
5.2. Типы СУБД и способы общения с ними..... 114
5.3. Язык семантических структур..... 117
5.3.1. Общие замечания..... 117
5.3.2. Запись запросов на языке SQL..... 118
5.3.3. Требования к семантической структуре (СемС)... 119
5.3.4. Формальное определение СемС..... 121
5.3.4.1. Узлы СемС..... 121
5.3.4.2. Дуги СемС..... 122
5.3.4.3. Поддеревья..... 122
5.3.5. Примеры СемС..... 126
5.4. Правила семантического анализа..... 129
5.4.1. Нормализация СинтС..... 129
5.4.2. Семантизация нормализованной СинтС..... 134
5.4.2.1. Семантизация лексики..... 135
5.4.2.2. Семантизация синтаксических отношений..... 149
5.4.3. Канонизация СемС..... 158
5.4.4. Образцы правил семантического анализа..... 162
5.4.4.1. Общее правило нормализации..... 162
5.4.4.2. Трафаретное правило нормализации..... 163
5.4.4.3. Словарное правило нормализации..... 164
5.4.4.4. Общее правило семантизации..... 165
5.4.4.5. Трафаретное правило семантизации..... 166
5.4.4.6. Словарное правило семантизации..... 168
5.4.4.7. Общее правило канонизации..... 169
5.4.4.8. Трафаретное правило канонизации..... 170
5.4.4.9. Словарное правило канонизации..... 171
5.5. Алгоритм семантического анализа..... 172

Глава 6. Словари лингвистического процессора 173
6 1 Комбинаторный словарь и его место в процессе переработки текста 173
6 1 1 Типы информации о лексеме 173
6 1 2 Классификационная информация в словарной статье КС 174
6 1 2 1 Вход словарной статьи и описание многозначных слов 174
6 1 2 2 Зона части речи 176
6 1 2 3 Зона синтаксических признаков 177
6 1 2 4 Зона семантических признаков 205
6 1 2 5 Зона моделей управления 210
6 1 3 Операционная информация в словарной статье КС 214
6 1 3 1 Правила синтаксического анализа 215
6 1 3 2 Правила нормализации 215
6 1 3 3 Правила семантизации 217
6 1 4 Образцы словарных статей КС 218
6 2 Семантический словарь 219
6 2 1 Типы информации в семантическом словаре 219
6 2 2 Образцы словарных статей семантического словаря 223

Глава 7. Эксперименты по семантическому анализу и машинному переводу 225
7 1 Эксперименты по семантическому анализу 225
7 2 Эксперименты по русско-английскому МП 232
7 2 1 Вводные замечания 232
7 2 2 Основные параметры двунаправленной системы машинного перевода ЭТАП-3 234
7 2 3 Новые свойства системы 236
7 2 4 Примеры работы русско-английской версии системы ЭТАП-3 237
7 2 4 1 Текст статьи об электронном оборудовании 238
7 2 4 2 Запросы к демонстрационной базе данных ORACLE 242
7 2 4 3 Экспериментальные предложения 244

Заключение 247

Литература 250

Научное издание

АПРЕСЯН Юрий Дереникович
БОГУСЛАВСКИЙ Игорь Михайлович
ИОМДИН Леонид Лейбович и др

ЛИНГВИСТИЧЕСКИЙ ПРОЦЕССОР
ДЛЯ СЛОЖНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

Утверждено к печати Институтом проблем передачи информации РАН

Редактор Т И Охотникова Художник А Г Кобрин
Художественный редактор В Ю Яковлев Технический редактор М Н Комарова

ИБ № 49237

Подписано к печати 10 05 92
Формат 60 X 90 1/16 Бумага офсетная № 1 Печать офсетная
Усл печ л 16 Усл кр отт 16,дУч, изд л 14 9
Тираж 400 экз Тип Зак 1538

Ордена Трудового Красного Знамени издательство Наука
117864 ГСП 7 Москва В-485 Профсоюзная ул 90

Ордена Трудового Красного Знамени
Первая типография издательства Наука
199034 Санкт Петербург В 34 9 линия д 12

Оригинал макет подготовлен на компьютере в Институте проблем передачи информации РАН