1
71
21

ИНСТИТУТ РУССКОГО ЯЗЫКА АН СССР

Проблемная группа по экспериментальной
и прикладной лингвистике

ПРЕДВАРИТЕЛЬНЫЕ ПУБЛИКАЦИИ

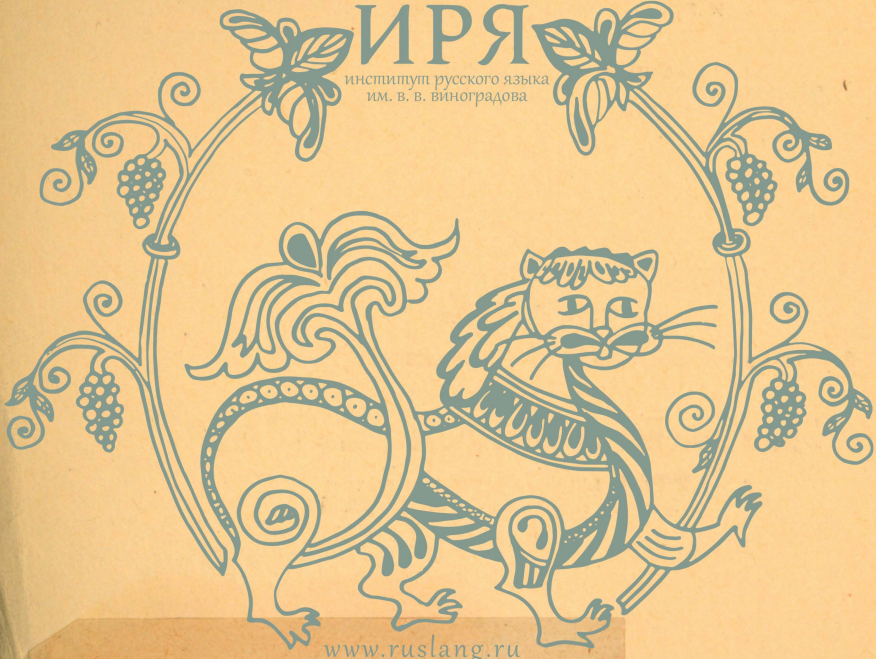
Выпуск 21

О. С. КУЛАГИНА, И. А. МЕЛЬЧУК,
К. О. ЭРАСТОВ

ОБ ОДНОЙ ВОЗМОЖНОЙ СИСТЕМЕ МАШИННОГО ПЕРЕВОДА

Москва — 1971

ИРЯ
институт русского языка
им. в. в. виноградова



www.ruslang.ru

66112

ИНСТИТУТ РУССКОГО ЯЗЫКА АН СССР
Проблемная группа по экспериментальной
и прикладной лингвистике

Предварительные публикации

Выпуск 21

О.С.Кулагина, И.А.Мельчук, К.О.Эрастов

ОБ ОДНОЙ ВОЗМОЖНОЙ СИСТЕМЕ МАШИННОГО ПЕРЕВОДА

Москва - 1971

П-721
БИБЛИОТЕКА
ИН-ТА РУССКОГО ЯЗЫКА
ИН-ТА ЯЗЫКОЗНАНИЯ АН СССР
66112

Если подходить к машинному, или автоматическому, переводу не как к чисто теоретической проблеме, а как к практической задаче, то есть видеть конечную цель исследований в создании действующей системы перевода, то необходимо параллельно с принципиальными, "стратегическими" вопросами обдумывать и сопряженные с ними возможные конкретные "технические" решения. Пренебрежение любым из этих двух аспектов обычно заводит работу в тупик: при недостаточном внимании к теоретической стороне проблемы создаются действующие, но весьма несовершенные и не допускающие усовершенствования системы; другая крайность - системы, глубоко продуманные, в идеале достаточно "мощные", но практически нереализуемые.

Опыт исследований в области машинного перевода за истекшие два десятилетия подсказывает, что наиболее реальный путь к созданию действующей системы, обеспечивающей достаточно качественный перевод и допускающей усовершенствования, проходит где-то посредине между однобокой приземленностью и однобокой теоретичностью и предполагает постоянную компромиссную увязку принципиальных и практических моментов в процессе разработки системы. Этот процесс, таким образом, мыслится как серия последовательных приближений.

Описываемая в данной работе лингвистическая концепция некоторой системы автоматического перевода (АП) представляет собой попытку продвижения по такому пути. Ее можно рассматривать как опыт конкретизации общей программы действий по созданию системы АП с учетом новейших теоретических и экспериментальных результатов в данной области, опубликованной в конце 1969 года. (Кулагина-Мельчук-Розенцвейг 1969). В соответствии со сказанным выше о желательном подходе к разработке систем перевода, предлагаемые принципиальные решения воплощены в достаточно конкретных схемах отдельных компонентов системы, разработанных с большей или меньшей детальностью. Это описание - больше чем просто иллюстрация, хотя его никоим образом нельзя рассматривать как нечто законченное. Вынесенный в заглавие

эпитет "возможный" следует понимать не в условном смысле принципиальной возможности, а как прямое утверждение о реальной осуществимости системы именно в том виде, в каком она здесь обрисована. Для удобства изложения мы часто говорим о системе в настоящем времени, т.е. как бы о существующей, но читателю следует все время иметь в виду, что речь идет о некотором проекте, и притом только в его лингвистическом аспекте.

Над этим проектом работал научный коллектив, в который, помимо авторов настоящей публикации, входили сотрудники Лаборатории машинного перевода ИТПИИЯ им. М.Тореза: В.З.Розенцвейг; Н.В.Перцов, И.И.Убин, А.П.Чехов и З.М.Шалашина, а также Н.Г.Арсентьева (ИПМ АН СССР).

I. Основные особенности системы АРАП.

1) В системе АРАП предполагается использовать большое количество лингвистических сведений об обоих привлекаемых языках, в частности - об индивидуальных особенностях английских и русских слов. Средний объем информации об отдельном слове, отдельной синтаксической конструкции и т.п. должен быть существенно больше того, что, насколько нам известно, имеется в других системах АП. Эти сведения сосредоточены, в основном, в словаре, и их использование должно способствовать преодолению различных известных трудностей, возникающих при АП (выбор правильных переводных эквивалентов, учет лексической сочетаемости и т.п.). Важно отметить, что при этом нужные свойства лексических единиц характеризуются по стандартной схеме, принятие которой предполагает введение ряда новых средств описания поведения слов. К числу таких средств в частности относятся:

- модели управления, детально описывающие синтаксическую сочетаемость (в частности, "сильное управление") каждого английского и русского слова (см. ниже, II, стр.18, раздел 321);

- лексические функции - такие достаточно элементарные "смыслы", которые обычно имеют и в английском и в русском языке фразеологически связанное оформление, зависящее от того

слова, при котором данный смысл должен быть выражен*).

Фиксация лексических функций в словаре и их учет при обработке текста уменьшает трудности, связанные с выбором нужного эквивалента. В то же время использование лексических функций и моделей управления помогает при установлении синтаксических связей, при выборе нужной связи среди нескольких альтернативных и т.д. (см. ниже, II, §§ 3,4).

*) Например, смысл "очень" (лексическая функция Magn) выражается при слове задача словом - трудная, при необходимость - настоятельная, при доказывать - убедительно, с несомненностью, как дважды два. Введение лексических функций и фиксация их в словаре позволяет удобным и системным способом учесть и описать многочисленные случаи полисемии, как, например, в следующем случае: paper I (бумага) vs. paper II (статья): в сочетаниях (1) interesting paper, (2) a blotting paper, (3) paper (written) by NN, (4) a sheet of paper слова interesting и (written) by являются значениями лексических функций, соответственно, Bon (смысл 'хорошая') и Caus (смысл 'создавать') от paper II 'статья', а слова sheet и blotting - значениями, соответственно, лексической функции S$_{ing}$ (смысл 'частица') и индивидуальной (особой) лексической функции (смысл 'промокательная'), имеющейся при paper I. В тексте слово paper, встретившееся в сочетаниях (1) и (3), будет, благодаря наличию в его словарной статье указанных значений лексических функций, правильно опознано и переведено как статья (интересная статья; статья (написанная) NN), а в сочетаниях (2) и (4) - как бумага (соответственно, промокательная бумага и лист бумаги).

Таким же путем будет обеспечен выбор правильного значения и у других многозначных слов в тех же примерах (sheet - лист, а не простыня; blotting - промокательная, а не капающая). Интересно, что даже такие близкие значения слова paper как 'статья' и 'доклад' тоже могут быть различимы с помощью лексических функций: to deliver (=лексическая функция Oper$_1$) a paper - 'читать, делать, доклад'; collection (= функция Mult) of papers - 'сборник статей' и т.д.

Подробнее о лексических функциях см. Холковский-Мельчук 1967г.

2) В истории автоматического перевода можно условно выделить три "поколения" систем АП.

К первому поколению (пятидесятые годы) относятся системы, в которых делалась установка на пословный или почти пословный перевод (с частичным использованием синтаксического анализа как вспомогательного средства для улучшения пословного перевода).

Системы второго поколения (шестидесятые годы) ставили во главу угла синтаксический анализ и перевод на уровне синтаксических структур. Главной целью при этом было получить правильное описание синтаксического строения переводимой фразы. Что же касается ее смыслового анализа, без которого правильный перевод часто бывает невозможен, то в этом отношении в системах второго поколения не было сделано существенного шага вперед по сравнению с предыдущим этапом.

Систему АРАП можно отнести к третьему поколению систем АП (системы семидесятых годов). Здесь предполагается осуществлять перевод через некоторое представление фразы, более близкое к ее инвариантному смысловому представлению, чем обычная поверхностно-синтаксическая структура, а именно через глубинно-синтаксические структуры. Более глубокое - "истинно" смысловое - представление пока не может быть использовано в АП, поскольку даже само содержание понятия "смысловое представление" далеко не выяснено, не говоря уже о соответствующем формализме. С другой стороны, менее глубокое представление - поверхностно-синтаксическая структура - слишком часто оказывается неконгруэнтным для пары "английский - русский"*).

*) Вообще говоря, чем менее схожи между собой какие-либо два языка, тем глубинней и ближе к чисто смысловой записи должно быть то представление фразы входного языка, через которое осуществляется переход к эквивалентной ей по смыслу фразе выходного языка. Сходство или несходство языков определяется при этом такими факторами, как совпадение/несовпадение порядка слов, наличие во входном языке синтаксических конструкций, невозможных для выходного языка (как, например, complex object в английском), и т.д. В предлагаемой системе имеется в виду обеспечить такую "глубинность" представления текста, которая соответствовала бы степени несходства между английским и русским языками.

Использование в качестве "мостика" для перевода с английского языка на русский специального уровня представления текста – уровня глубинно-синтаксических структур, промежуточного между чисто синтаксическим и чисто семантическим уровнями, составляет вторую особенность системы АРАП. Следует отметить, что в этом отношении предлагаемая система имеет большое сходство с Гренобльской системой французско-русского АП (Мельчук 1969).

Углубление в смысл переводимого текста естественным образом связано с увеличением размера рассматриваемого при переводе отрезка текста. Все существующие в настоящее время системы АП рассчитаны на перевод одной фразы, рассматриваемой изолированно от остальных фраз того же текста; межфразовые связи в них никак не учтены. Ясно, однако, что при решении ряда вопросов (установление антецедентов местоимений, разбор некоторых случаев многозначности и т.п.) учет контекста более широкого, чем одна фраза, может оказаться необходимым. В системе АРАП предполагается сохранять некоторые данные о переведенной фразе с тем, чтобы они могли быть использованы при переводе последующих фраз (на первых порах – одной последующей фразы для решения упомянутых выше вопросов; пределы учитываемого контекста будут расширяться по мере разработки проблем межфразовых связей).

3) Третья особенность описываемой системы состоит в том, что в ней все этапы синтеза должны быть разработаны гораздо детальнее, чем это обычно делалось в существующих системах АП.

До сих пор в центре внимания, как правило, находился анализ, синтез же считался операцией практически однозначной и потому простой. В результате нередко случалось, что достаточно богатая информация, извлеченная алгоритмом анализа из входного текста, не могла быть в полной мере использована одноцелевым*) и максимально стандартизованным алгоритмом синтеза. Это приводило к тому, что в выходных фразах нередко оказывалось большое количество неправильностей и шероховатостей.

*) Т.е. рассчитанным на выдачу одного основного варианта.

каждая из таких нероховатостей не так уж опасна сама по себе, однако их нагромождение в одной фразе или в соседних фразах может существенно ухудшать текст (что и имеет место в переводах даже Гренобльской системы – на наш взгляд, одной из лучших среди существующих в настоящее время в мире).

В системе АРАП синтез задуман как достаточно сложная операция, обеспечивающая гибкое идиоматичное выражение всей информации, извлеченной при анализе.

При таком подходе синтез, естественно, оказывается многоцелевым (многовариантным) и по сложности не уступает анализу. В частности, многие операции, которые традиционно включаются в анализ, например, выбор наиболее подходящего в данном контексте переводного эквивалента, в тех случаях, когда этот выбор зависит не от того, в каком из своих значений взято английское слово*), перемещены в синтез за счет использования аппарата лексических функций. При таком подходе исследователям удается решать многие вопросы, не выходя за пределы родного языка, что представляет ряд преимуществ.

Такое усиление внимания к синтезу связано также и с проблемой многозначности. Дело в том, что для выражения некоторого заданного смысла на данном языке достаточно, грубо говоря, хорошего владения данным языком. В то же время, для того, чтобы извлечь смысл из текста на данном языке владения данным языком в общем случае заведомо недостаточно: обычно

*) Число таких случаев чрезвычайно велико. Так, например, слово launch имеет около двадцати русских переводов, которые с точки зрения английского языка, соответствуют одному словарному значению этого слова (launch a satellit – запустить спутник, launch a ship –спустить на воду корабль, launch a proposal – выдвинуть предложение и т.д.); соответствующие русские эквиваленты в предлагаемой системе будут извлекаться на этапе синтеза из русской части словаря – как записанные в словарных статьях (слов спутник, корабль и т.д.) значения лексической функции $CausFunc_0$.

требуется понимание того, о чем идет речь, масса сведений о реальном мире, умение делать логические выводы и т.п. Именно этим обусловлено существование омонимии (в широком смысле слова), неразрешимой не только для машины, но и для человека, не являющегося специалистом в данной области знания. Таким образом, анализ и синтез в АП неравноправны.

Поскольку ввод в машину такого количества сведений, которое позволило бы ей правильно понимать смысл любой фразы, в настоящее время неосуществим, избран следующий путь. Во всех случаях, где английская фраза допускает - с точки зрения машины - более одного понимания, система АРАП должна обнаружить все возможные понимания. В тех случаях, когда для выбора правильного понимания достаточно только сведений о данном языке, система должна обеспечить разрешение омонимии. Если же одних только сведений о языке для этого недостаточно и для правильного понимания фразы необходимы некоторые сведения о действительности и т.п., то от системы требуется только, чтобы каждое из альтернативных пониманий было выражено на выходе идеале безупречным русским текстом. При этом некоторое понимание может быть и неправильным; однако если переводящая фраза сохраняет его смысл и вполне читабельна, то одноязычный (русский) постредактор-специалист, следящий за смыслом текста в целом и контролирующий работу системы, легко сможет через русский перевод браковать выбранное "английское" понимание и требовать от системы другого. Таким образом, недостижимый во многих случаях автоматический выбор при анализе одного правильного варианта среди многих, равновозможных для машины, перекладывается на плечи человека. Однако задачу этого человека мы стремимся максимально упростить за счет развитого синтеза: постредактору придется иметь дело лишь с хорошим русским текстом и судить о его осмысленности.

4) Ввиду значительной сложности системы АРАП оказалось необходимым расчленить ее на компоненты, для каждого из которых максимально полно и четко описаны вход и выход. Благодаря этому достигается независимость отдельных компонентов, внутреннее строение которых можно менять, лишь бы это не затра-

гивало их входов и выходов, которыми они только и связаны с соседними компонентами. Это облегчает разработку, реализацию и экспериментальную проверку отдельных компонентов.

Входы и выходы отдельных компонентов системы АРАП суть уровни представления текста, соответствующие уровням языка принятым в лингвистике; тем самым каждый из этих уровней является содержательно обоснованным, благодаря чему оказывается возможным независимое исследование и описание лингвистами таких уровней безотносительно к системе. Процесс перевода представляет собой последовательность преобразований представления одного уровня в соответствующее представление другого уровня (см. в Приложении пример последовательных представлений фразы на разных уровнях в процессе ее перевода)[*].

5) В силу отмеченной выше сложности системы АРАП и очевидной необходимости ее непрерывного усовершенствования в процессе эксплуатации система перевода в собственном смысле должна быть включена в систему программ обслуживания. Подчеркнем, что здесь имеются в виду не технические программы обслуживания, а по существу программы, в какой-то степени автоматизирующие исследовательскую работу создателей системы. Система обслуживания должна решать задачи трех типов:

---

*) Отметим, что многоуровневые модели довольно широко употребляются в настоящее время для структурных описаний языка (ср. работы P.Sgall'a, S.Lamb'a и др.); характерно, что в последнем вышедшем в США вузовском учебнике "вычислительной лингвистики" многоуровневым моделям посвящен специальный параграф (Hays 1967: 161-169). Предлагаемое блочное строение системы АРАП представляет собой, таким образом, возможный способ технического воплощения этой плодотворной идеи.

(а) Внесение всех необходимых исправлений и дополнений в систему перевода (расширение словаря, уточнение или добавление правил и т.п.). При этом должно автоматически проверяться, что внесение изменений не приводит к непредусмотренной "порче" системы в каком-либо другом месте.

(б) Выдача по запросам сведений о лингвистических и иных данных, входящих в систему. Примерный характер запроса: выдать все словарные статьи с такими-то признаками; выдать все правила синтаксического компонента системы с совпадающими левыми (или правыми) частями и т.п.

(в) Регулирование режима работы системы. Предполагается, что в процессе отладки, а также эксплуатации системы АРАП может понадобиться временное отключение тех или иных ее компонентов, переключение того или иного компонента с полного перебора вариантов на выдачу первого подходящего варианта, настройка на текст определенной тематики*) и т.п. Кроме того, система должна быть снабжена механизмом, позволяющим учитывать внесенные постредактором поправки при продолжении перевода (см. выше, стр. 9).

Многие из идей, на которых основывается система АРАП, высказывались ранее: некоторые из них были воплощены в действующих системах перевода. Главное новшество заключается в объединении всех упомянутых идей и одновременном и последовательном их проведении в рамках одной системы.

Предлагаемая система достаточно сложна и ее реализация потребует значительных усилий. Однако опыт работы над автоматическим переводом в нашей стране и за рубежом показывает, что системы, имеющие существенно более простое строение, перевод хорошего качества дать не могут.

*) Подобная настройка облегчает автоматическое снятие многозначности в случаях типа calculus = 1) исчисление [математика, логика], 2) камень [почечный и т.п., медицина].

Итак, особенности системы АРАП, определяющие ее строение, таковы:

1) Использование обширной лингвистической информации, в особенности, словарного характера; особая роль словаря (лексические функции, модель управления и т.п.).

2) Перевод через глубинно-синтаксический уровень.

3) Повышенное внимание к синтезу.

4) Многоуровневость: выделение компонентов с четко определенными входами и выходами.

5) Установка на максимальную автоматизацию процедур усовершенствования системы*).

II. Словарь системы АРАП.

Словарь системы представляет собой совокупность словарных статей, каждая из которых содержит максимально полное описание всех свойств одной английской или русской лексической единицы (ЛЕ). Поскольку лексическая единица - это слово (лексема) или несвободное словосочетание (фразема), взятое в одном значении, каждому выделяемому значению любого слова или несвободного словосочетания соответствует своя словарная статья.

Для удобства словарь разделен на английскую и русскую части; обе эти части связаны между собой взаимными отсылками (в каждой английской статье указаны заглавия словарных статей русских переводных эквивалентов, а в каждой русской словарной статье даются отсылки к тем английским словарным статьям, для которых данная русская ЛЕ является переводным эквивалентом).

Каждая из названных частей словаря - английская и русская - подразделяется, в свою очередь, на словарь лексем (т.е. основ) и словарь фразем.

---

*) В общем виде идея предлагаемой системы была изложена в работе Kulagina-Mel'čuk, 1967.

Все словарные статьи, как русские, так и английские, как для лексем, так и для фразем имеют однотипное строение, различаясь несущественным образом в соответствии с различием в их содержательной природе. Поэтому мы опишем здесь строение словарной статьи только для английских лексем.

Словарная статья состоит из следующих пяти зон (которые делятся на разделы, подразделы и т.д.).

0) Заглавная лексическая единица (ЛЕ).

1) Морфологические сведения (часть речи, словоизменение и регулярное словообразование).

2) Семантические сведения (сведения о лексической производности заглавной ЛЕ, ее русский переводной эквивалент и (семантические признаки).

3) Синтаксические сведения (синтаксический код, модель управления и дополнительные условия к ней).

4) Лексические сведения (значения лексических функций).

Ввиду той центральной роли, которую, как уже отмечалось, должен играть словарь в системе АРАП, его строение описано с большой подробностью. Читатель, не интересующийся частностями организации словаря, может без ущерба для понимания пропустить остальную часть раздела П, с тем, чтобы читая о работе отдельных блоков системы, вернуться, там, где он почувствует в этом потребность, к описанию зон и разделов словарной статьи, используемых соответствующим блоком*).

Зона 0. Заглавная лексическая единица (ЛЕ).

0I. Заглавная лексическая единица, которую характеризует данная словарная статья: например, UNIT 1(2), где цифра в скобках показывает общее число значений, различаемых у многозначной буквенной "оболочки" данной ЛЕ, а первая цифра - номер значения, характеризуемого данной словарной статьей; при ЛЕ с однозначной буквенной "оболочкой" таких

*) Образцы словарных статей (английских и русских) см. в работе: Образцы 1970.

цифр, естественно, нет. Заглавная ЛЕ записывается в виде основы; если у данной ЛЕ имеются супплетивные основы (образующиеся нестандартным образом) или графические варианты, такие основы и варианты указываются в подразделе ОИ (например, для лексической единицы BE в подразделе ОИ будет записано IS, ARE, WAS, WERE; для ЛЕ CENTER в подразделе ОИ будет записано CENTRE*).

Если заглавная ЛЕ является фраземой, то в зоне ОИ записывается либо последовательность основ, образующих фразему, - FOR INSTANCE, POINT OF DEPARTURE, ONCE (AND) AGAIN, IN ALL PROBABILITY (для случаев, когда компоненты фраземы не могут разделяться другими словами), либо глубинно-синтаксическое дерево** фраземы - BUILD $\xrightarrow{5}$ UP, TAKE $\xrightarrow{2}$ PLACE, DRIVE $\xrightarrow{3}$ OF $\xrightarrow{2}$ AIR $\xrightarrow{5}$ THE 'глушить, забивать' [радар] (для случая разрывных фразем).

Зона 1. Морфологические сведения

11. Часть речи заглавной ЛЕ: S - существительные (в том числе субстантивные местоимения и количественные числительные), V - глаголы, A - прилагательные, Adv - наречия, Inv - неизменяемые слова (предлоги, союзы и др.). Каждой ЛЕ можно приписать только один символ части речи.

12. Словоизменение: указание на способ получения всех словоизменительных форм от данной основы. Это делается за-

---

*) Подобные основы (супплетивные и вариантные) выступают в качестве заглавных ЛЕ отдельных словарных статей - так называемых неполных, в которых дается только отсылка к основной статье (и морфологическая характеристика, если надо). Так, FEET имеет особую статью, где указано только, что это мн.число (раздел 23, см. стр. 17) от FOOT ; аналогично, словарная статья PROGRAMME содержит только отсылку к PROGRAM : и т.п.

**) О глубинно-синтаксической структуре см. ниже, стр. 39 и сл.

данием словоизменительного класса, или типа парадигмы (подраздел 121), и словоизменительных особенностей (подраздел 122), если таковые имеются; особенности указываются путем перечисления тех словоизменительных форм, которые от данной ЛЕ образовать нельзя.

Пример заполнения разделов 11 и 12 для основы CAN 'мочь':

11. V

121. Класс 4

122. нет {INF}, {ING} , {PART}

(то есть от CAN 'мочь' нельзя образовать инфинитив и причастия I и II).

13. Словообразование: указание на способы получения производных основ от данной основы. Например, в словарной статье PRODUCE:

13.{ION}, {IVE},

где символы в фигурных скобках обозначают словообразовательные морфемы, которым соответствуют в конкретных случаях различные словообразовательные суффиксы - морфы.

В данном разделе учитывается только такое словообразование, которое является регулярным ("правильным") и строго "аддитивным" (значение производной основы = значение исходной основы + значение аффикса): read + able ⟹ readable, move + ment ⟹ movement, consider+ation ⟹ consideration

Тем самым, регулярные и строго аддитивные производные могут, вообще говоря, не заноситься в словарь в качестве заглавных ЛЕ. Разумеется, что при нарушении формальной регулярности или смысловой аддитивности производные должны иметь свои отдельные словарные статьи (например, consideration в значении 'соображение', ср. Из соображений экономии и т.п.).

14. Тип чередований, которые могут происходить у данной основы при словоизменении и словообразовании. Пример: для основы PRODUCE:

14. A_{-ce-ct} | R+{ION}, {IVE},

то есть "Когда от данной основы R образуется существительное или прилагательное посредством указанных морфем, исход основы PRODUCE, т.е. -ce, переходит в -ct"; в результате

образуются основы PRODUCTION, PRODUCTIVE.

Сведения, записанные в зоне 1, используются правилами блока "Английская морфология" при анализе, и блока "Русская морфология" при синтезе (см. ниже, §§2, 10).

Зона 2. Семантические сведения.

211. Лексическая производность: здесь указывается та лексическая функция (ЛФ) = замена, значением которой является данная заглавная ЛЕ; ясно, что этот раздел заполняется не у всех ЛЕ, а только у ЛЕ, являющихся значением некоторой лексической функции = замены. (Сведения этого раздела являются, в известном смысле, обратными к сведениям, записываемым в разделе 41.). Примеры: COOK2 [повар] = S_1(COOK1), ACHIEVEMENT = S_{res}(ACHIEVE), SOLAR=S_0(SUN).

212. Переводной эквивалент. В большинстве случаев каждая английская ЛЕ (= одно значение!) имеет один русский эквивалент. В тех случаях, когда переводных эквивалентов более одного, при них могут быть даны условия выбора, например:

ЛЕ PRECISELY, раздел 212:

1. РОВНО | G(C_0) - количество
2. ИМЕННО | G(C_0) - не количество,

т.е. перевод ровно выбирается, если слово precisely ($=C_0$) зависит от слова с признаком "количество" (precisely five lines), а перевод именно - в противном случае. Если условия выбора указать невозможно, то будут выдаваться все имеющиеся эквиваленты.

213. Семантические признаки: не более десяти признаков типа "лицо", "процесс", "информация", "одушевленность", образующих либо булеву формулу, либо синтаксическое дерево; всего разных семантических признаков предполагается иметь не более 200. Семантические признаки учитываются при проверке допустимости тех или иных сочетаний ЛЕ; см., в частности, стр. 19 о разделе 322.

22. Примеры употребления заглавной ЛЕ.

Сведения данного раздела в машину не вводятся; они служат лишь для удобства человека, работающего со словарем.

23. Грамматическая информация, которая не может быть получена при морфологическом анализе (см. ниже § 2), то есть сведения о нестандартно образуемых формах; например, для английской ЛЕ FOOT будут заведены отдельные статьи FOOT и FEET, и в разделе 23 статьи FOOT будет указание о единственном числе, а в статье FEET – о множественном.

Зона 3. Синтаксические сведения.

31. Общие синтактико-комбинаторные возможности заглавной ЛЕ: описываются указанием синтаксических признаков – подклассов слов, типа "существительное – имя собственное", "предлог направления" – и сведений о синтаксических конструкциях, в которые может входить данная ЛЕ. (Более подробной характеристики этого раздела авторы дать пока не могут в связи с неразработанностью блоков синтаксического анализа, см. ниже, стр. 35.)

32. Сведения о синтаксическом оформлении слов, заполняющих так называемые смысловые валентности ("места") данной ЛЕ, то есть связанных с ЛЕ теми семантико-синтаксическими (глубинно-синтаксическими – см. ниже, § 4) связями, которые регулярно – обязательно или достаточно часто – выражаются при данной ЛЕ; таких типов связи различается шесть – четыре "актантных" ("предикат – субъект", "предикат – 1-й, 2-й, 3-й объект"), одна определительная (в широком смысле), и одна сочинительная (между однородными членами). В данном разделе записываются требования и условия к словам, заполняющим четыре актантных места (их семантические признаки, синтактико-морфологическое оформление, совместность/несовместность, т.е. сочетаемость тех или иных заполнений разных мест). Сведения, помещаемые в разделе 32, используются при анализе как в процессе перехода от фразы к ее поверхностно-синтаксическому представлению, так и при последующем переходе к ее глубинно-синтаксическому представлению, а при синтезе русской фразы – в процессе обратных переходов: от глубинно-синтаксического к поверхностно-синтаксическому представлению, а затем – к выходному тексту.

В соответствии с той важной ролью, которую играет информация раздела 32, он имеет сравнительно большой объем и дробную структуру.

66112
БИБЛИОТЕКА
ИН-ТА РУССКОГО ЯЗЫКА
ИН-ТА ЯЗЫКОЗНАНИЯ АН СССР

321. Модель управления, фиксирующая синтактико-морфологические требования к словам, зависящим от заглавной ЛЕ по глубинно-синтаксическим связям 1-4.

Пример: модель управления ЛЕ METHOD.

М1	М2	М3
кто пользуется методом	для чего применяется метод	в чем состоит метод
1. S_{poss} 2. $Pron_{poss}$ 3. of+S	1. for+S 2. of+S 3. of+V_{ing} 4. S_{attr} *)	1. S_{attr} 2. $V_{ing,attr}$ 3. $A_{o}(S)$

Эта запись называет синтактико-семантические валентности слова method и указывает альтернативные способы заполнения каждой из них, а именно:

- субъект пользования методом (М1) может выражаться или притяжательной формой существительного (John's method of doing this), притяжательным местоимением (our method of doing this) или именной группой с предлогом of (The method of John for doing this was ...);

- назначение метода (М2) выражается любым из четырех способов: существительным с предлогом for (method for computation of the test), существительным с предлогом of (method of computation ...), существительным в "атрибутивной" позиции (the computation method), а также герундием (the method of computing);

- наконец, приемы или процедуры, из которых складывается метод (М3), выражаются тремя способами: "атрибутивным" существительным (approximation method), герундием (approximating method), а также прилагательным (approximative method).

Одним из способов оформления мест в модели управления может быть, в частности, придаточное того или иного типа; так, в графе М2 модели управления ЛЕ FACT2 имеем:

*) Символ S_{attr} обозначает существительное, стоящее непосредственно перед данным в конструкции типа stone wall.

that + ПРЕДЛ

(т.е., например: we draw reader's attention to the fact that the prime requirement is said to be ...)

Некоторые ЛЕ могут иметь более одной модели управления, например, BELIEVE: He believes that B is C и He believes B to be C; каждому из проиллюстрированных здесь употреблений глагола believe отвечает своя модель управления (одна - двухместная, друга.. - трехместная). В таких случаях мы говорим о модификациях (вариантах) моделей управления.

322. Семантические требования к заполнению мест: семантические признаки слов, которые могут заполнять то или иное место при заглавной ЛЕ. Например, для заглавной ЛЕ СНЕСК1(4) в этом подразделе записано: (М2 : S = S ("информация" или "предикат" или "свойство"), что означает: "Слово, зависящее от данной ЛЕ по 2-му типу глубинно-синтаксической связи, должно быть существительным, обладающим хотя бы одним из перечисленных семантических признаков".

323. Требования к обязательности и сочетаемости мест: сведения о возможном/невозможном или обязательном одновременном выражении тех или иных мест заглавной ЛЕ теми или иными способами. Эти сведения задаются формулами следующего типа:

для ЛЕ ORDER2:М3.1 и М4.1 несовместимы, т.е. "При глаголе order (заказывать) невозможно одновременно два беспредложных существительных в функции 'исполнитель заказа' и 'получатель заказа' соответственно: либо I ordered the tailor (=М3.1) clothes for my father, либо I ordered my father (М4.1) clothes at the tailor's, но не *I ordered my father clothes the tailor".

для ЛЕ TELL : М2 или М3 обязательно, то есть "У глагола tell обязательно заполнение хотя бы одного места - второго или третьего" (tell somebody или tell something, но нельзя просто tell).

Зона 4. Лексические сведения.

41. Значения стандартных лексических функций. Сведения, содержащиеся в данном разделе, являются опорными для работы блоков "Английский глубинный синтаксис" и "Английское перифразирование", см. ниже, §§ 4 и 5.

В русской части словаря сведения этого раздела используются, соответственно, блоками "Русское перифразирование" и "Русский глубинный синтаксис" - §§ 7, 8.

Поскольку лексические функции являются принципиальным новшеством предлагаемой системы, кратких пояснений в рамках описания структуры словарной статьи недостаточно для того, чтобы охарактеризовать их полностью, и поэтому здесь мы ограничимся примером; ср. сноску к стр. 5.

Пример заполнения раздела 4I в словарной статье ERROR:

$Magn_1$: bad; serious
$Magn_2$: large
$AntiMagn_2$: small
$Magn_3$: flagrant
$Oper_1$: commit, make, introduce | D_1=S("одуш"или"прибор")*)
Liqu : correct

Соответственно, в русской части словаря в разделе 4I статьи слова ОШИБКА дано:

$Magn_1$: опасная; серьезная
$Magn_2$: грубая
$AntiMagn_2$: незначительная, несущественная
$Magn_3$: вопиющая
$Oper_1$: делать, допускать ~ У|D_1=S("одуш"или"прибор")
Liqu : исправлять, устранять ~ У

(вертикальной чертой отделены условия, при которых данная ЛФ имеет указанное значение).

42. Индивидуальные лексические функции: это ЛФ, определенные только для небольшого числа аргументов. Специальных обозначений для них не вводится; выбор правильного перевода выполняется для них следующим образом: все индивидуальные ЛФ некоторой английской ЛЕ перенумерованы; соответствующие индивидуальные ЛФ русского эквивалента этой ЛЕ получают те же номера. Например, среди переводных эквивалентов слова

*) Т.е. первое место глагола $Oper_1$ должно быть заполнено существительным с признаком "одушевленность", или "прибор".

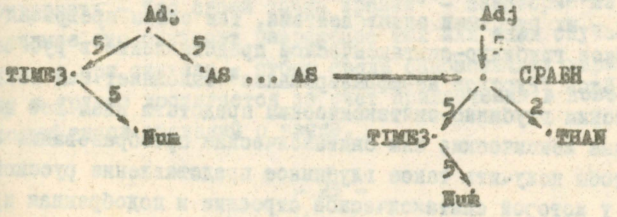
explanatory русского толковый не будет; правильный перевод сочетания explanatory dictionary = толковый словарь обеспечивается переходом через соответствующую индивидуальную ЛФ, указанную в зоне 42:

В английской части словаря:	В русской части словаря:
DICTIONARY	СЛОВАРЬ
F_1 : explanatory	F_1 : толковый

43. Фразеология: здесь указываются те фраземы, в которых данное слово является ключевым (т.е. фраземы, сильно прогнозируемые данным словом). При анализе на основе этих указаний выполняется обращение к соответствующим фраземам и обнаруженье их в тексте. (Более подробно охарактеризовать здесь использование данного раздела невозможно, поскольку в целях экономии места мы не описали типы фразем, способы записи их в словаре и т.п.).

44. Индивидуальные преобразования: связанные с данной ЛЕ нестандартные преобразования структуры, отличающиеся от общих правил перифразирования (которые выполняются согласно сведениям разделов 321 и 322); эти преобразования нужны для приведения английской глубинно-синтаксической структуры к так называемому базовому виду, облегчающему переход к русской структуре (см. ниже, § 5). Например, специфически английский оборот three times as much as (=в три раза больше) должен быть заменен синонимичным ему базовым three times more than, который по своей структуре ближе к соответствующему русскому обороту.

Для этого в данном разделе словарной статьи TIMES (= раз) будет записано правило:



где Adj - прилагательное, Num - числительное, а СРАВН - фиктивная ЛЕ (стр. 40) в роли показателя сравнительной степени.

III. "Грамматика" системы АРАП.

§ 1. Состав "грамматики".

Грамматика системы АРАП состоит из девяти блоков, каждый из которых отвечает за переход от одного уровня представления фразы к другому. 4 блока работают с английским текстом в процессе анализа, 4 - с русским (в ходе синтеза); 1 блок выполняет преобразование результатов анализа в исходные данные для синтеза. [см. стр. 27].

Перечислим эти блоки.

1. Блок "Английская морфология" - совокупность правил, которая (вместе с английской частью словаря) обеспечивает морфологический анализ словоформ входного текста.

2. Блок "Английский поверхностный синтаксис" - совокупность правил, которая обеспечивает получение для входной фразы ее поверхностно-синтаксического представления.

3. Блок "Английский глубинный синтаксис" - совокупность правил, позволяющая переработать поверхностно-синтаксическое представление в глубинно-синтаксическое.

4. Блок "Английское перифразирование" - совокупность правил, позволяющая перейти от глубинно-синтаксического представления входной фразы к ее базовому представлению.

5. Блок "Выбор русских эквивалентов", который получая на вход глубинное представление рабочей фразы (безразлично-базовое или нет), заменяет английские лексемы - с помощью словаря - их русскими эквивалентами, тем самым превращая английское глубинно-синтаксическое представление в русское.

6. Блок "Русское перифразирование" выполняет над полученным русским глубинно-синтаксическим представлением все необходимые лексические или синтаксические преобразования с тем, чтобы получить такое глубинное представление русской фразы, у которой синтаксическое строение и подобранная предыдущим блоком лексика "согласованы" друг с другом.

7. Блок "Русский глубинный синтаксис" – перерабатывает глубинно-синтаксическое представление выходной русской фразы в поверхностно-синтаксическое.

8. Блок "Русский поверхностный синтаксис" – перерабатывает поверхностно-синтаксическое представление в упорядоченную цепочку символов всех лексем выходной фразы (со всеми необходимыми морфологическими характеристиками и знаками препинания). При этом выполняются следующие задачи:

а) определение полного набора морфологических характеристик для выходных русских лексем;

б) определение линейного порядка словоформ выходной фразы;

в) обработка повторяющихся слов: замена их местоимениями (он, она, оно, они, себя, свой ...) и опущение (эллипсис);

г) расстановка знаков препинания.

9. Блок "Русская морфология" – превращает символ лексемы со всеми ее морфологическими характеристиками в реальную словоформу (в принятой орфографической записи).

Каждый из перечисленных здесь блоков является в принципе многовариантным. Формально это значит, что в общем случае одному "входу" каждый блок может сопоставлять несколько альтернативных "выходов", равноправных с точки зрения этого блока, т.е. при том объеме информации, которой он располагает. Содержательно же это соответствует:

при анализе-наличию неразрешимой (средствами данного блока) омонимии, а при синтезе-наличию синонимичных вариантов, выбор между которыми в пределах данного блока невозможен.

Альтернативные результаты каждого блока поступают на вход следующего блока, который может в процессе переработки отвергать – "со своей точки зрения" – некоторые из них. При анализе это означает разрешение тех или иных случаев омонимии, а при синтезе – отбор среди синонимичных вариантов; и то и другое достигается за счет привлечения в новом блоке новых типов сведений о языке.

Подчеркнем, что для простоты изложения мы не будем в дальнейшем оговаривать тот факт, что некоторым данным, поступившим на вход любого блока, может не соответствовать никаких выходных результатов.

Среди перечисленных блоков особые надежды возлагаются на блоки 4–6 "Английское перифразирование" + "Выбор русских эквивалентов" + "Русское перифразирование". Именно здесь осуществляется переход от английского языка к русскому; в блоках 5 и 6 подбираются все те русские языковые средства, которые должны позволить последующим блокам синтеза построить хороший русский текст для любого из полученных при анализе пониманий входной фразы, в том числе – и для неправильных пониманий неоднозначных фраз. Выше – стр. 9 – уже говорилось, что это должно помочь однозначному постредактору или читателю выбрать нужное понимание и в случаях не снятой омонимии.

Что же касается самого разрешения омонимии в ходе анализа, то для этого в системе АРАП никакого специального блока не предусматривается. Сведения, обычно используемые для снятия омонимии, достаточно разнородны – от морфологических и синтаксических до лексических и семантических, и поэтому представляется более целесообразным "разнести" процесс снятия омонимии по соответствующим этапам обработки фразы.

Так, лексико-грамматическая омонимия должна полностью различаться в блоке поверхностно-синтаксического анализа (кроме тех случаев, когда вся фраза в целом синтаксически неоднозначна). Например, в обеих фразах:

(1) They lead there 'Они ведут туда'

(2) Their lead is there 'Их свинец там'

слово lead после морфологического анализа получит две характеристики – как существительное и как глагол, но после синтаксического анализа в первой фразе будут вычеркнуты сведения о существительном, а во второй – о глаголе. Однако, во фразе

(3) Lead pigs 'Свинцовые чушки' и
'Ведите свиней',

синтаксический анализ разумеется, не поможет, так как фраза в целом не однозначна.

Что касается лексической омонимии, то иногда она может сниматься при синтаксическом анализе, но может остаться и неразрешенной. Снятие лексической омонимии в блоке глубинно-синтаксического анализа будет обеспечено в тех случаях, когда омонимичные ЛЕ имеют разные модели управления (§ 1, раздел 32И).

Пример:

SET 2 ПОМЕЩАТЬ

М1 [субъект]	М2 [объект]	М3 [характеристика объекта]
S	S	Loc_{ad}+S

SET 3 НАСТРАИВАТЬ, РЕГУЛИРОВАТЬ

М1 [субъект]	М2 [объект]	М3 [цель регулировки]
S	S	1. to+V_{inf} 2. for+S

Различие в моделях управления данных с..ов обеспечит снятие омонимии, например, в следующих фразах:

(1) The device is set to respond quickly to a signal. = Устройство настраивается таким образом, чтобы оно быстро реагировало на сигнал.

(2) The value is set into the register. = Величина помещается в регистр.

Таким образом, в каждом из блоков анализа могут разрешаться определенные случаи омонимии, не снятой в предыдущем блоке. Кроме того, в блоке "Выбор русских эквивалентов" будет осуществляться разрешение "квази-омонимии" - выбор одного из альтернативных русских эквивалентов для слова, с точки зрения английского языка имеющего одно значение:

например, wash 'делать чистым с помощью жидкости' соответствует русским 'стирать/мыть', go = 'итти/ехать', bring = 'принести/привести' и т.п. Этот выбор будет осуществляться на основе сведений о сочетаемости русских слов, записанных в их словарных статьях: (в разделах 32 - "Модель управления", 41 - "Лексические функции", и 421 - "Индивидуальные лексические функции". Так, например, слову WASH в словаре сопоставлено два русских эквивалента мыть и стирать ; в подразделе 322 русской словарной статьи МЫТЬ указано, что 2-е место (объект действия) не может быть заполнено словом с семантическим признаком "-ткань", при слове же СТИРАТЬ имеется обратное указание. Благодаря этому, WASH в контекстах wash the floor и wash the table-cloth будет, соответственно, переведено МЫТЬ (пол) и СТИРАТЬ (скатерть). Однако такой проясняющий контекст может отсутствовать, и тогда система выдаст два разных перевода.

Каждый блок содержит совокупность правил вида $\alpha \Longrightarrow \beta$, где α и β - это соответствующие друг другу элементарные "куски" представлений двух смежных уровней. Все правила выписываются как бы в направлении синтеза, т.е. α - это всегда элемент более глубинного (=более близкого к смыслу) представления; при этом правила группируются по совпадающим левым частям. Такой способ записи оказывается более обозримым и удобным для лингвистов; кроме того при этом правила анализа и синтеза имеют одинаковое строение (хотя при анализе правила должны, разумеется, применяться в "обратном" направлении).

Этапы преобразования информации в системе АРАП

Синтез

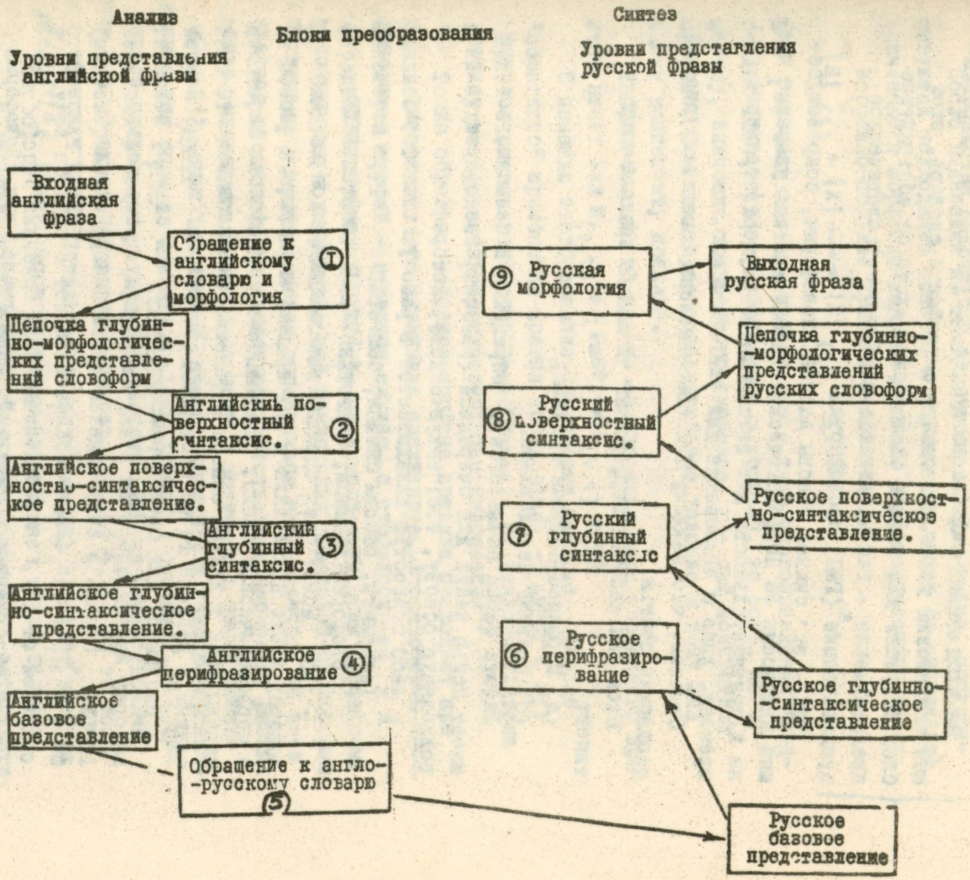
Уровни представления русской фразы

Блоки преобразования

Анализ

Уровни представления английской фразы

ЭТАПЫ ПРЕОБРАЗОВАНИЯ ИНФОРМАЦИИ В СИСТЕМЕ АРАП



§ 2. Английская морфология.

На вход данного блока поступает побуквенно закодированный английский текст. Используя английский словарь, этот блок получает *для каждой словоформы* текста ее абстрактное представление - так называемое глубинно-морфологическое представление*) (ГМП) словоформы.

ГМП - это символ основы данной словоформы, сопровождаемый индексами её морфологических характеристик: указание числа и притяжательности для английского существительного, времени, лица и наклонения для глагола и т.п.

Отметим, что каждому ГМП сопоставлена соответствующая словарная статья.

Блок "Английская морфология" включает правила двух типов:

(а) Морфемно-морфные правила;

(б) Правило чередования.

Работа блока начинается с обращения к словарю, которое имеет целью найти для очередной (рабочей) словоформы текста подходящую основу, т.е. такую основу, которая:

- либо полностью вкладывается в рабочую словоформу (например, основы COMMON, COMPUTER - при текстовых словоформах COMMON, COMPUTERS *или* COMPUTERIZE;
- либо может быть сделана вкладывающейся в рабочую словоформу после применения к ней правил чередования, имеющихся в морфологическом блоке; применимость того или иного правила чередования к данной основе определяется по сведениям, записанным в разделе 14 её словарной статьи (см. выше, §1).

Например, для словоформы applied в словаре нет основы, которая вкладывалась бы в эту словоформу: есть только основа APPLY. Однако в силу наличия правила чередования Cy + C ⟶ Ci + C | C ≠ i ("Буква y после согласной перед гласной, отличной от i, заменяется на стыке морф буквой i") основа appl оказывается подходящей. Аналогично, для словоформы

*) Термин "глубинно-морфологическое представление" (словоформы) противопоставляется термину "поверхностно-морфологическое представление", которым обозначается изображение словоформы в виде цепочки морфем.

taking словарная основа TAKE оказывается подходящей благодаря правилу e + $\Gamma \longrightarrow$ + Γ и т.д.

После выделения из текстовой словоформы части, соответствующей подходящей основе, к оставшейся части (к окончанию) применяются морфемно-морфные правила морфологического блока.

Морфемно-морфные правила имеют вид

$$\{R\}_{(i)} + \{X\} \Longrightarrow \{R\}_{(i)} + \underline{x} \mid C,$$

где {R}-основа, i - информация о типе словоизменения (подраздел 121), {X} - суффиксальная морфема, то есть, по существу, значение конкретного суффикса, $\underline{x}$ - сам конкретный суффикс, C - некоторое дополнительное условие к тому или иному компоненту правила.

Например, одно из правил образования форм прошедшего времени имеет вид $R_{(v,1)} + \{PAST\} \Longrightarrow R_{(v,1)} + ed$.

С помощью этого правила словоформа applied (после "обратного" применения правил чередования - APPLY + ed) будет опознана как форма Past (Tense) от данного глагола и ей будет приписана соответствующая информация*).

Таким образом, морфемно-морфные правила позволяют при анализе отождествить окончание анализируемой словоформы с некоторой морфой - в данном случае, с суффиксом -ed - и определить морфему ({PAST}), представляемую этим суффиксом.

Следует подчеркнуть, что разложение текстовой словоформы на основу и суффикс может выполняться неоднозначно, то есть некоторым словоформам может соответствовать несколько разных ГМП. Эта неоднозначность связана с наличием в языке лексической (seals 'печати' и seals 'тюлени') и лексико-грамматической (lead 'свинец' и lead 'вести') омонимии. В соответствии с общей установкой системы АРАП на множественный (многовариантный) анализ и синтез (ср. §§ 4, 6, 7) во всех

*) Другое правило с такой же правой частью припишет этой же словоформе значение Past Participle, и таким образом ее морфологический анализ даст неоднозначность интерпретации, которая в нормальном случае будет снята на этапе поверхностно-синтаксического анализа (см. ниже, § 3).

подобных случаях морфологический блок будет выдавать все возможные разложения словоформ на основу и суффикс и, соответственно, несколько ГМП.

§ 3. Английский поверхностный синтаксис.

На вход данного блока поступает цепочка ГМП (или альтернативные цепочки ГМП), полученные из фразы морфологическим блоком.

Результатом работы синтаксического блока является поверхностно-синтаксическое представление (ПСП) фразы, т.е. следующая совокупность сведений:

1) Поверхностно-синтаксическая структура (ПСС) - поверхностное дерево зависимостей рабочей фразы.

2) Указания об анафорических связях для местоимений, союзных слов типа which и т.д.

3) Указания о синтаксических группах (о составляющих).

4) Сведения о логических акцентах.

Пример поверхностно-синтаксического представления фразы приведен в Приложении, рис. 1.

Подчеркнем, что важнейшим из этих четырех компонентов ПСП является дерево ПСС, которое прежде всего подвергается обработке и перестройкам на всех последующих этапах, превращаясь в глубинно-синтаксическую структуру, которая затем сводится к базисной, подвергается перифразированию и т.д.; остальные группы сведений играют скорее вспомогательную роль (ср. ниже §§ 4, 5, 6). Ввиду этого, мы в последующем изложении часто говорим о структуре (ПСС, ГСС), а не о представлении (ПСП, ГСП), хотя фактически на всех этапах представления фраза содержит, кроме дерева, и остальные элементы информации.

Охарактеризуем подробнее названные выше компоненты ПСП.

1) Поверхностное дерево зависимостей представляет собой дерево, в узлах которого стоят ГМП, но без тех морфологических признаков, которые не несут смысловой информации, а являются чисто согласовательными: это притяжательный ("саксонский родительный") падеж существительного, признак "косвенная форма" у личных местоимений, лицо и число у глагола. В противополож-

ность этим признакам число существительного, время и наклонения личного глагола обязательно сохраняются при узлах ПСС. Ветвям этого дерева сопоставлены типы поверхностно-синтаксических связей (отношения непосредственной доминации – ОНД); всего предполагается различать порядка 100 таких типов связей.

Примеры ОНД:

Предикативное	he ←предик.— said
1-е комплетивное	saw —1 компл.→ him
2,3,4-ое комплетивное	bought —2 комп.→ him, —3 компл.→ a book, —4 компл.→ from N for $1
Определительное	special ←опред.— type
Отпредложное	by —отпредл.→ Shakespeare
Обстоятельственное	went —обстоят.→ there
Служебное	shall —служеб.→ argue
Однороднсе	A —однород.→ and —однород.→ B
Композитное	stone ←композ.— wall

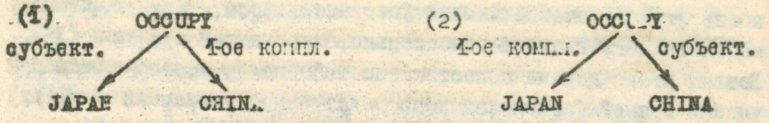
В дереве зависимостей (после окончания синтаксического анализа) сведения о линейном порядке узлов, т.е. о порядке слов во фразе не сохраняются. Вся нужная информация, которая передается порядком слов, должна в результате синтаксического анализа быть представлена явным и однозначным способом либо при помощи ОНД, либо при помощи других компонентов ПСП.

Например, для сочетания (1) stone wall "каменная стена" и (2) wall stone "стенной камень" – "камень стены", где порядок слов определяет, какое из них главное, а какое зависимое, в ПСП будет получено разное направления стрелок композитного ОНД:

(1) WALL ←композ.— STONE (=stone wall)

(2) WALL —композ.→ STONE (=wall stone)

Для фраз (1) Japan occupied China "Япония оккупировала Китай" и (2) China occupied Japan "Китай оккупировал Японию", отличающихся только порядком слов, будут получены разные ПСС:



т.е. информация, передаваемая в тексте порядком слов, будет в явном виде выражена наименованием соответствующего ОНД.

2) Указания об анафорических связях: для местоимений-заменителей (it, he, they, these) и союзных слов типа which указывается их антецедент - существительное. Это делается для того, чтобы при переходе к английской глубинной структуре заменить все подобные слова их антецедентами. При переходе к русской фразе местоимения вводятся в русский текст независимо, по собственным правилам русского языка (дело в том, что законы введения местоимений в русском и английском языках существенно различаются, и прямая передача английских местоимений русскими может вести к серьезным ошибкам).

Пример: во фразе It would be possible to operate the system at less than its full capabilities слово its прямо перевести невозможно (its = её? его? их?); сначала необходимо установить анафорическую связь между its и system, затем подставить в ГСС system вместо its (получив что-то вроде at less than full capabilities of the system), далее получить перевод - Можно было бы запустить систему, используя меньше, чем полные возможности системы и, тогда наконец, можно ввести правильное притяжательное местоимение её: ... используя меньше, чем её полные возможности.

Еще один пример:

(1) As soon as the first two subroutines computer the results, the third subroutine prints these out.

(2) As soon as the first two subroutines compute the results, the third subroutine prints them out.

Согласно законам английского языка, местоимение these будет опознано блоком синтаксического анализа как заменитель существительного the results, а местоимение them - как заменитель существительного subroutines (местоимения соеди-

нены с антецедентами пунктирными стрелками). В русском языке передать эту разницу местоимениями нельзя, и для получения однозначных переводов необходимо повторить в тексте антецедент местоимения, который нужно для этого предварительно знать из анализа английской фразы. Переводы:

(1) Как только первые две подпрограммы вычислят результаты, третья подпрограмма выпечатывает эти результаты.

(2) Как только первые две подпрограммы вычислят результаты, третья подпрограмма выпечатывает эти подпрограммы.

Заметим, что поскольку антецедент местоимения может находиться в предшествующей фразе, так что анафорические связи - в отличие от собственно синтаксических - могут связывать словоформы из двух смежных фраз.

3) Указание о синтаксических группах означает использование, в дополнение к дереву зависимостей, некоторых сведений о структуре составляющих фразы. Эти сведения используются эпизодически в тех случаях, когда это необходимо для однозначного анализа. Например, для однозначного анализа (и перевода) словосочетания old men and women недостаточно знать структуру соответствующего куска дерева зависимостей, отражающего наличие однородной связи между существительными men и women, а необходимо установить также, относится ли определение old только к первому существительному (перевод в таком случае старики и женщины) или к обоим (перевод: старики и старухи), то есть определить структуру составляющих. В первом случае структура такова:

(old men) and women,

во втором

old (men and women).

В подобных случаях однозначно определить структуру составляющих на этапе поверхностного анализа не всегда возможно. Существенно поэтому учесть оба альтернативных разложения на составляющие с тем, чтобы в дальнейшем либо выбрать одно из них при опоре на более широкий контекст, может быть, с выходом за пределы фразы; либо, когда контекст не помогает, попытаться воспроизвести ту же неоднозначность на

этапе русского синтеза или же оставить оба понимания на выбор человеку, по возможности отметив то из них, которое представляется наиболее вероятным с точки зрения нормального строения английской фразы.

4) Сведения о логических акцентах, или указание, что именно является в данной фразе темой (о чем говорится), а что - ремой (что именно утверждается), необходимы в связи с тем, что противопоставление темы и ремы существенным образом сказывается на порядке слов в русской фразе, тогда как в английском языке это противопоставление выражается достаточно сложным способом с помощью весьма разнородных средств (артикли, cleft sentences - фразы типа It is the above device that does the transformation , порядок слов и связанный с ним выбор активного или пассивного залога и др.). Поскольку трудно непосредственно соотнести эти средства с нужным русским порядком слов, решено в ходе анализа вырабатывать указания о теме и реме. В тех случаях, когда включенных в алгоритм правил окажется недостаточно для обнаружения темы и ремы, можно считать, что темой является группа грамматического подлежащего, а ремой - группа сказуемого.

Поясним это на следующем примере:

(1) It was yesterday that this letter was mailed;

(2) It was this letter that was mailed yesterday.

Выделение ремы в этих фразах осуществлено в виде cleft sentence с помощью оборота It was ... that. При анализе этот оборот вызовет появление в полученном ПСП указания о реме (подчеркнутые слова), которое будет реализовано в русских переводах с помощью соответствующей расстановки слов:

(1) Это письмо отправлено вчера.

(2) Вчера отправлено это письмо.

(Ср. также пример в сноске на стр. 59).

Для получения по входной английской фразе (точнее - по цепочке глубинно-морфологических представлений ее словоформ) охарактеризованного выше ПСП данный блок использует описание английского синтаксиса. Это описание будет включать средства, достаточно разнообразные по содержательной природе.

Для построения поверхностного дерева зависимостей (=ПСС) предполагается использовать средства двух типов:

А) Синтагматические сведения (микросинтаксис),

Б) Фразовые сведения (макросинтаксис).

А) Синтагматические сведения.

К синтагматическим сведениям относятся: описания синтактико-комбинаторных возможностей лексических единиц (см. § 4 раздел 31), сведения об обязательных валентностях ЛЕ (также содержащиеся в их словарных статьях, см. § 4, раздел 321), а также перечень поверхностно-синтаксических правил. Поверхностно-синтаксическое правило - это выражение вида

$$T \Longrightarrow X \mid C,$$

где T - некоторое поверхностное дерево зависимостей (как правило, пара узлов и связывающее их ОНД), X - цепочка ГМП (глубинно-морфологических представлений словоформ), отвечающая этому дереву, а C - различные дополнительные условия как на T, так и на X (компонент C может быть пуст). В цепочке X специально указывается контактное (A + B) или дистантное (A + ... + B) расположение соответствующих словоформ в тексте, а также их текстовый порядок; в тех случаях, когда порядок может быть любым, применяется запись A+B/B+A(A + ... +B/ /B + ... + A соответственно). Приведем примеры поверхностно-синтаксических правил (для большей наглядности - в упрощенном виде):

S_{sg}, ONE_{sg}
| артиклевое $\Longrightarrow$ A+...+ $\begin{Bmatrix} S_{sg} \\ ONE_{sg} \end{Bmatrix}$ | Между A и S_{sg}, ONE_{sg} могут стоять только элементы, подчиненные S_{sg} или ONE_{sg}
A

(a book, an interesting book, a very interesting book; a good one)

S,ONE
| артиклевое $\Longrightarrow$ THE+...+ $\begin{Bmatrix} S \\ ONE \end{Bmatrix}$ | Между THE и S,ONE могут стоять только элементы, подчиненные S или ONE
THE

(the book/books; the good ones; the one I saw)

S^1 : S^2 композитное $\Longrightarrow S^2+S^1$ | не: X · детермин · S^2

(information processing, systems theory research)

V : S предикативное $\Longrightarrow S+\ldots+V$ | Согласование S и V в числе

(The boy runs/ The boys run; The police sais/say; Method I and method IV are described...)

$V^{\equiv}$: S предикативное $\Longrightarrow V^{\equiv}+\ldots+S$ | Согласование S и V в числе; $V^{\equiv}$=do, have, be, can, shall, should, will, would, ...; в вопросительном предложении; между $V^{\equiv}$ и S могут стоять только элементы, подчиненные S.

(Has he come? Does this method, as has been noted, provide a means for that?)

Условия, применяемые в правилах, весьма различны по характеру; они могут касаться:

- наличия/отсутствия определенных признаков у отдельных элементов правила, а также соответствий между ними (сюда, например, относятся требования к согласованию сказуемого с подлежащим в двух последних правилах);
- наличия/отсутствия у элементов правила определенных слуг или хозяев ("несовместимость или неотделимость синтагм");
- допустимости тех или иных цепочек между элементами правой части правила (в случае их дистантного расположения);
- некоторых особенностей предложения (например, вопросительность).

Б) Фразовые сведения.

Фразовые сведения касаются строения английской фразы в целом. Фразовые сведения достаточно разнородные по своей природе, являются в некотором смысле господствующими: именно или

регулируется употребление синтагматических сведений. Можно указать следующие основные категории фразовых сведений:

а) Все, что относится к принятым нами условностям изображения синтаксической структуры, например: каждое слово, кроме вершины фразы, должно иметь одного хозяина ; расстановка стрелок отношений должна соответствовать поверхностно-синтаксическим правилам - контактному/дистантному расположению связанных слов, требованию проективности и т.д.

б) Правила предпочтения.

В процессе анализа часто возникают конфликтные ситуации: например, когда одно и то же слово можно связать с двумя потенциальными хозяевами в соответствии с разными поверхностно-синтаксическими правилами, Для подобных случаев формулируются правила предпочтения, определяющие иерархию удовлетворения требований; среди них будут, в частности, следующие правила:

1. Требование модели управления сильнее поверхностно-синтаксического правила.

2. "Функциональное" требование, т.е. связь между словом-значением лексической функции и ключевым словом (аргументом этой функции) сильнее поверхностно-синтаксического правила.

Так, при анализе фразы

I read a paper by Jones.

предлог by может быть интерпретирован как слуга глагола read (согласно правилу $V \xrightarrow{\text{обстоят.}} Pr \Rightarrow V+...Pr/Pr+...+V$) или существительного paper - согласно его модели управления (раздел 32I словарной статьи этого слова). Оба эти потенциальные хозяина могут подчинять предлог by, но соответствующее правило предпочтения требует удовлетворить в первую очередь валентность, указанную в модели управления. (Перевод фразы, соответственно, - Я читаю статью Джоунза).

Для фрагмента фразы

... lecture in mathematics delivered at 5 p.m.,

согласно правилу, правая часть которого $S+...+V_{pp}$ пассивное причастие delivered может быть связано определительным

ОНД как с первым, так и со вторым существительным (lecture → delivered или mathematics → delivered). Однако, в качестве хозяина для delivered будет выбрано слово lecture, поскольку, как указано в его словарной статье (раздел 4I), deliver является значением лексической фун.ции Oper$_1$ от lecture,

в) Правила несовместимости или обязательной одновременности слуг одного хозяина; по существу, это закономерности взаимодействия поверхностно-синтаксических правил, носящие весьма общий характер и потому не включенные в индивидуальные условия к поверхностно-синтаксическим правилам.

г) Правила обоснованности знаков препинания.

Учитывая, что правила пунктуации в английском языке недостаточно унифицированы, причем нередко в реальных текстах не соблюдаются даже общепринятые правила, синтаксический анализ не может опираться на пунктуацию как на одно из основных средств определения границ - сплошных или разрывных - фрагментов фразы, к которым применимы поверхностно-синтаксические правила), а также при установлении сочинительных связей и т.п. Тем не менее, знаки препинания непременно будут учитываться: при построении синтаксической структуры необходимо проверить, нет ли во фразе знаков препинания, противоречащих рассматриваемой структуре,
ср. *The man, I have met today, ...
атрибутивная

д) Частные правила типа: "В английском языке не бывает препозитивных распространенных определительных оборотов" (недопустимы сочетания типа *a rich in information book вместо a book rich in information и т.п.).

Широко известно, что одной из трудностей синтаксического анализа является многочисленность возможных структур для анализируемой фразы, т.е. распространенная в естественных языках синтаксическая неоднозначность. В соответствии с общим принципом системы АРАП блок синтаксического анализа также должен получать все возможные ПСП рабочей фразы. При этом не предполагается, что для фразы сначала будут строиться

все возможные альтернативные синтаксические структуры, а уж затем из них с помощью каких-либо других средств, например, семантики, будут выбираться нужные. В системе АРАП синтаксический анализ задуман так, чтобы при нем одновременно и в тесном взаимодействии друг с другом использовались как сведения о синтаксическом строении фраз английского языка (список поверхностно-синтаксических правил), так и сведения о лексической и семантической сочетаемости английских слов и фразем (информация синтаксической и лексической зон словаря).

Это, по-видимому, позволит, сохраняя принцип многовариантного анализа, получать для фразы лишь те поверхностно-синтаксические структуры, которые совместимы с данной лексикой и не противоречат семантической сочетаемости слов, учтенной в системе АРАП.

Что касается средств получения остальных компонентов ПСП, то в силу их неразработанности, мы не можем охарактеризовать их здесь сколько-нибудь подробно. Отметим только, что правила нахождения антецедентов местоимений явно должны будут выходить за пределы рабочей фразы и обращаться к предшествующей фразе.

§ 4. Английский глубинный синтаксис.

На вход данного блока поступает ПСП рабочей фразы (или несколько альтернативных ПСП, если используемых средств различения синтаксической омонимии оказалось недостаточно для выбора одного варианта).

На выходе блока выдается глубинно-синтаксическое представление фразы, отвечающее данному ПСП (см. Приложение, рис. 2). Глубинно-синтаксическое представление (ГСП) есть следующая совокупность сведений:

1) Глубинно-синтаксическая структура (ГСС) - глубинное дерево зависимостей рабочей фразы.
2) Указания о тождестве денотатов тех или иных именных групп.
3) Указания о синтаксических группах.
4) Сведения о логических акцентах.

1) Глубинное дерево зависимостей представляет собой дерево, в узлах которого стоят символы четырех типов:

а) Символы лексем - заглавия словарных статей, сопровождаемые самими словарными статьями.

б) Символы фразем, также сопровождаемые их словарными статьями.

в) Символы лексических функций (поскольку слова, являющиеся значением лексических функций от присутствующих в глубинном дереве и связанных с ними единиц типа а) и б), заменяются символами соответствующих ЛФ; например, если в ПСП имелось сочетание

write —(Ie комплетивное)→ program,

то в глубинном дереве слово write, являющееся - в соответствии со сведениями, записанными в зоне A статьи PROGRAM, - значением ЛФ Caus от program, исчезнет, а в соответствующем узле дерева вместо него появится символ Caus.

г) Символы фиктивных лексем (например, вводимый в ГСС сравнительных конструкций фиктивный символ СРАВН, см. выше, стр. 22).

Все эти символы сопровождаются индексами морфологических характеристик (в том числе - от аналитических форм), несущих семантическую информацию: число у существительных, время, наклонение, залог, аспект и перфектность у глагола. (Как было сказано на стр. 30, морфологическая информация, играющая чисто синтаксическую роль, была опущена еще на поверхностно-синтаксическом уровне).

Ветвям глубинного дерева сопоставлены типы глубинно-синтаксических связей, которых различается шесть: отношения 1–4 связывают имя предиката с именами его первого, ..., четвертого аргументов, соответственно; отношение 5 связывает определяемое и определение (при этом определение понимается в широком смысле); отношение 6 соединяет однородные члены.

Подчеркнем, что глубинно-синтаксические связи не имеют однозначного соответствия с поверхностно-синтаксическими связями (ОНД). Они отражают скорее смысловые отношения между синтаксически связанными словами, и с этой точки зрения их можно было бы назвать синтактико-семантическими связя-

ми *.

Так, например, атрибутивному ОНД может соответствовать как "определительная" глубинная связь (связь 5), так и одна из "актантных" (1 – 4) :

Поверхностное представление:

(1) British. определит isles (Британские острова)

(2) British. определит intervention (британская интервенция)

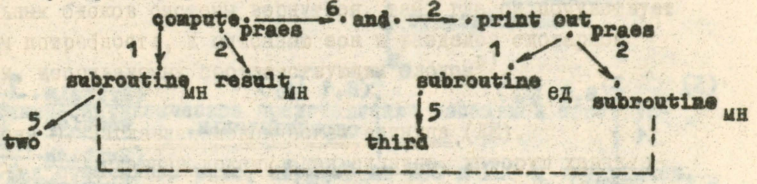
Глубинное представление:

(1) Britain. —5→ . isle мн

(2) Britain. —1→ . intervention

2) Поскольку все местоимения-заменители поверхностно-синтаксической структуры при переходе к ГСП заменяются их антецедентами, а также, поскольку здесь происходит восстановление некоторых типов эллипсиса (см. пример на стр. 45), то в ГСП необходимо специальным образом отмечать тот факт, что две разные именные группы обозначают один и тот же объект действительности (см. § 3).

Для этого фиксируется тождество денотатов таких групп, как это обозначено пунктирной линией на следующей схеме (фраза 2, стр. 32):



3) Указание о синтаксических группах (составляющих) и

4) Сведения о логических акцентах (тема-рема) сохраняются из ПСП (см. выше, § 3).

Преобразование ПСС в ГСС выполняется при помощи правил, которые имеют вид: $t_1 \longrightarrow t_2 \mid C$, где t_1 и t_2 – деревья зависимостей, соответственно, поверхностное и глубинное, а C – дополнительные условия к t_1 и t_2 (как и в прочих случаях,

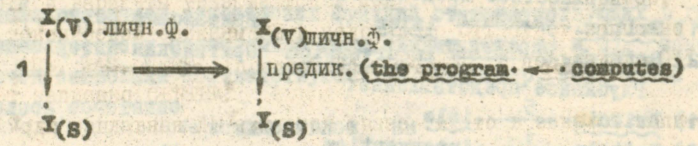
---

*) Правда, тот факт, что знаменательные слова сохраняются в узлах глубинного дерева целиком, а не разлагаются на смысловые компоненты, как требовал бы полный семантический анализ, в известной мере налагает на строение ГСС отпечаток поверхностной структуры, из которой эта ГСС получена.

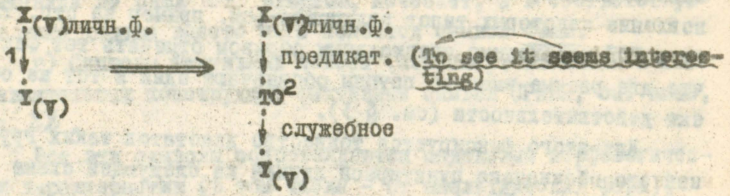
компонент C может быть пуст). Применение подобных правил состоит в том, что в ПСС отыскивается вхождение t_1 и оно заменяется на t_2 *).

Примеры правил глубинно-поверхностных преобразований **)

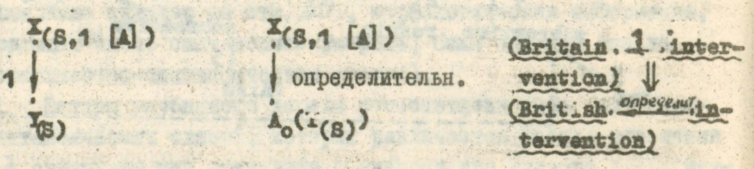
(1)



(2)



(3)

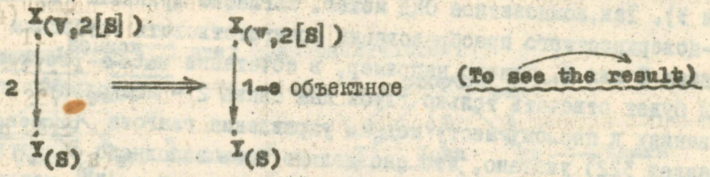


---

*) О правилах преобразований деревьев см. подробнее в работах: Кулагина 1969, Гладкий-Мельчук 1971.

**) Как и выше (в разделе, посвященном блоку "Английская морфология"), правила преобразования выписаны здесь в сторону синтеза; см. стр. 16 .

("Глубинное отношение 1 между У и У может соответствовать определительному ОНД между Х и относительным прилагательным, образованным от У, если Х - это существительное, у которого 1-ое место может быть выражено прилагательным, а У - существительное").

(4) 

("Глубинное отношение 2 между Х и У может соответствовать 1-ому объектному ОНД между ними, если Х - это глагол, у которого 2-ое место может быть выражено беспредложным существительным, а У - существительное").

(5) 

(Глубинно-синтаксическое представление пассивной конструкции).

Как отмечалось выше, переход от ПСС к ГСС не однозначен, поскольку существует правила с совпадающими правыми частями *).

Таким образом, одной ПСС могут соответствовать несколько альтернативных ГСС, что отвечает общей установке системы на многовариантный анализ и синтез.

---

*) С другой стороны, несколько разных ОНД могут переходить в одну и ту же глубинную связь: ср. выше примеры правил преобразования (1) и (3).

Подчеркнем, однако, что в процессе преобразования поверхностных структур в глубинные некоторые варианты будут отфильтровываться (в частности, за счет модели управления с ее различными требованиями к глубинным местам - § 1, раздел 3). Так, композитное ОНД может, согласно правилам глубинно-поверхностного преобразования соответствовать глубинным связям 1 или 2. Однако, например, в сочетании data ←(комп.) processing ОНД будет отвечать только глубинная связь 2 - поскольку в требованиях к первому месту модели управления глагола process (раздел 322) указано, что оно должно быть заполнено существительным с семантическим признаком "мыслящее устройство"[3], напротив, для сочетания machine ←(комп.) processing будет получена глубинная связь 1, поскольку слово MACHINE имеет признак "мыслящее устройство", что отмечено в семантической зоне его словарной статьи.

Кроме того, будут отфильтровываться также варианты, противоречащие синтактико-комбинаторным возможностям лексических единиц (раздел 31 словарной статьи).

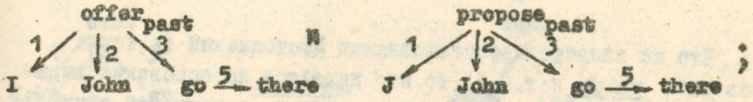
Важно отметить, что порядок применения правил преобразования поверхностных структур в глубинные лингвистически несущественен. Это означает, что правила должны быть составлены так, чтобы результат их применения никогда не зависел от порядка, в котором они применяются.

Применение описанных глубинно-поверхностных правил дает, так сказать, "предварительную", или "первичную", английскую ГСС; для ее дальнейшей переработки необходимо "восстановить" в ней ряд элементов, которые в реальном тексте опускаются или заменяются местоимением по законам английского языка, в результате чего получится "окончательная", или "приведенная" ГСС.

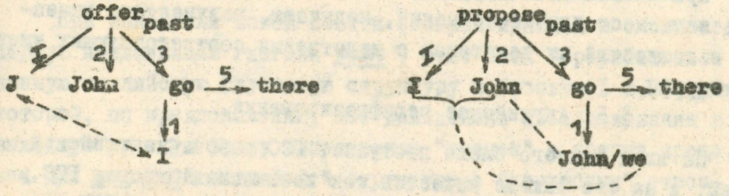
Так, для фраз I offered John to go there и I proposed John to go there "предварительные" ГСС будут сходными:

---

[3] "Мыслящее устройство" - человек, организация, наука, логический автомат и т.п.



"приведенные" же ГСС этих фраз различаются:



Подобное восстановление опущенных мест (в данном случае, у go) выполняется по специальным правилам, опирающимся на сведения в модели управления глаголов offer и propose; например, в словарной статье OFFER указано: М1 (М3) = М1, тогда как в словарной статье PROPOSE имеется другое указание: М1 (М3) = М2/М2 + М1.

Другим характерным примером восстановления в ГСС опущенных элементов является обработка конструкции с регулярным для английского языка употреблением двух "сочиненных" предлогов перед одним существительным, например:

Electric drive moves the device to and from the table.

В этой фразе блоком "Английский глубинный синтаксис" будет восстановлено пропущенное, в соответствии с нормами английского языка, второе вхождение слова table : в результате получится представление с повторяющимся словом (... moves to the table and from the table). Позже, уже на этапе русского поверхностного синтеза (см. ниже, § 9), это представление будет обработано правилами, которые обеспечат подстановку местоимения вместо второго (повторяющегося) вхождения слова столик в соответствии с нормами русского языка:

Электрический привод передвигает устройство к столику и от него (а не *... к и от столика).

Что же касается восстановления местоимений (а также слов типа which и т.п.), то это делается на основании выработанных ранее сведений о соотнесенности местоимений с их антецедентами, т.е. на основании анафорических связей, зафиксированных в ПСП.

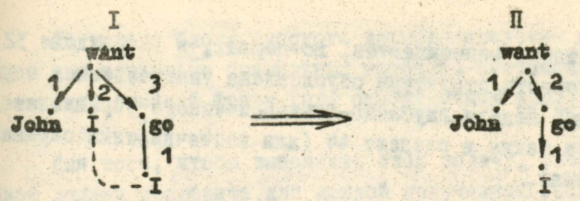
При этом именные группы, возникающие при восстановлении опущенных мест или местоимений, получают, разумеется, указания о тождестве их денотатов с денотатами соответствующих именных групп.

§ 5. Английское перифразирование.

На вход данного блока поступает ГСС рабочей английской фразы, а на его выходе выдается так называемая базовая ГСС. Базовая ГСС - это представитель группы синонимичных ГСС, выбираемая на основе критериев формальной простоты, максимальной эксплицитности и минимальной идиоматичности. В силу этих свойств именно для базовой ГСС наиболее вероятно иметь достаточно прямые соответствия в выходном языке. С точки зрения перевода это означает, что английские базовые ГСС удобно принять в качестве исходных данных для русского синтеза.

Отметим, однако, что приведение английских ГСС к базовому виду вовсе не является обязательным: во многих случаях вполне удовлетворительно переводятся на русский язык и не базовые ГСС. Поэтому данный блок пускается в ход только тогда, когда в английской ГСС имеются лексические единицы, снабженные специальными пометами о необходимости преобразования данной ГСС в базовую. Такими пометами снабжаются те лексические единицы (лексемы и фраземы), синтаксические свойства которых резко отличаются от синтаксических свойств их русских эквивалентов.

Например, из следующих двух синонимичных структур базовой считается вторая, а первая, непосредственно не переводимая на русский, должна получить помету о необходимости преобразования:



(John wants me to go) (John wants that I go)

При появлении такой пометы (сопровождающей в словаре первую модификацию глагола **want**) этот блок перерабатывает данную английскую глубинную структуру в базовую(I в II), для которой, по предположению, все дальнейшие преобразования в выходную русскую фразу выполняются "прямее" и потому проще, чем для исходной английской структуры с "неудобным" строением.

В действительности для целей перевода может быть достаточным локальное приведение к базовому виду, т.е. приведение к базовому виду не всей структуры, а лишь тех ее подструктур, в которых содержатся "неудобные" лексемы.

Данное преобразование выполняется правилами того же типа, что и правила, используемые в следующем блоке - "Русское перифразирование", поэтому мы не будем останавливаться здесь детально на виде правил (см. стр.49-50). Важно, однако, подчеркнуть, что организация применения правил в данном случае иная, чем в блоке "Русское перифразирование". Там задача состоит в построении синонимичных вариантов входной русской структуры до тех пор, пока очередной вариант - причем в принципе заранее не известно, какой именно, - не окажется удовлетворительным с точки зрения критериев, имеющихся в русских словарных статьях: требования к заполнению тех или иных мест модели управления, условия употребимости данной лексемы в данной конструкции и т.п. (см. ниже, стр.49). Здесь же цель иная: для варианта, содержащего специально отмеченные неудобные конструкции, нужно получить совершенно определенный базовый вариант. Перебора здесь не требуется, поскольку для каждой неудобной единицы явно указано, каким именно образом ее надо преобразовать для приведения к базовой форме.

Эти указания записываются, во-первых, в разделе 32 (когда небазовость структуры обусловлена употреблением одной из модификаций модели глубинно-синтаксического управления данной ЛЕ), а также в разделе 44 (для нестандартных случаев преобразования).

§ 6. Выбор русских эквивалентов.

На вход блока "Выбор русских эквивалентов" поступает ГСП рабочей английской фразы (с ГСС, приведенной, если это было нужно, к базовому виду блоком английского перифразирования).

На выходе этого блока получаются ГСП, где русские глубинно-синтаксические структуры получены из ГСС английской фразы заменой английских лексем их русскими эквивалентами, указанными в соответствующих словарных статьях; остальные компоненты ГСП сохраняются без изменения.

В результате подстановки русских эквивалентов получаются русские ГСС, которые могут быть как правильными, то есть такими, что их лексика и синтаксическое строение уже "согласованы" друг с другом, так и неправильными, то есть нуждающимися в дальнейшей перестройке, необходимой для обеспечения соответствия лексики и структуры. Эта перестройка осуществляется на следующем этапе работы системы - в блоке "Русское перифразирование".

Как уже говорилось выше, при выборе русских эквивалентов для лексических единиц, однозначных с точки зрения английского, но многозначных с точки зрения русского языка (имеющих более одного несинонимичных эквивалента), происходит разрешение случаев такой квази-омонимии (см. выше, стр. 25).

§ 7. Русское перифразирование.

На вход этого блока поступает русское ГСП. Его ГСС может быть как правильной, так и неправильной, с точки зрения семантических, синтаксических и прочих требований, записанных в словарных статьях русских слов.

На выходе блока русского перифразирования получается одно или несколько правильных глубинно-синтаксических представлений русской фразы (или синонимичных фраз). (См. Приложение, рис. 3).

Для того, чтобы выполнить свою задачу, рассматриваемый блок должен проверить для каждой полученной ГСС, удовлетворяет ли она требованиям, записанным в словарных статьях русских слов, стоящих в узлах, и в случае, если эти требования не выполняются, применить правила перифразирования с тем, чтобы в результате получить правильные русские ГСС *).

Для проверки полученной ГСС на правильность (на соответствие требованиям, записанным в словаре) для каждого самостоятельного русского слова **) надо найти его словарную статью в русском словаре, взять его модель управления ***) и проверить, выполняются ли синтактико-морфологические и семантические требования к заполнению мест, указанные в этой модели. Кроме того, надо проверить, выполняются ли требования к обязательности и сочетаемости мест, а также специальные синтактико-комбинаторные ограничения и требования нестандартных обязательных преобразований (раздел 44).

Если эти требования выполняются для всех узлов данной ГСС, то это значит, что структура правильна с точки зрения русского языка и правила перифразирования применять не нужно. Если же какое-то требование нарушено, то необходимо применить правила перифразирования.

Правила перифразирования состоят из семантических правил (правил синонимичных замен), сопровождаемых указанием соответствующих правил синтаксической перестройки, которую

---

*) Подробнее о процедуре синонимичных преобразований (перифразирования) см.: Арсентьева 1969, Жолковский-Мельчук 1967.

**) То есть, для лексических единиц, не являющихся значениями лексических функций.

***) Этот раздел русской словарной статьи имеет ту же структуру, что и соответствующий раздел английской – см. выше, § 1.

необходимо проделать над фразой в случае применения данного семантического правила. Такой перестройке подвергается обычно не вся фраза в целом, а только тот узел, в котором стоит заменяемая по семантическому правилу ЛЕ, и непосредственно подчиненные ему узлы.

Семантические правила представляют собой указания того, какой лексической функцией можно заменить некоторую лексическую функцию от обрабатываемого правилом слова (C_0), или само это слово (в некоторых случаях с указанием контекста, а в других – без такого указания), чтобы – после осуществления соответствующей синтаксической перестройки – смысл фразы остался без изменения.

Приведем пример применения правил перифразирования к русской ГСС, соответствующей английской фразе:

Power consumption of the rader is relatively high.

После обработки английской ГСС этой фразы блоком выбора русских эквивалентов будет получена русская ГСС, соответствующая переводу:

Потребление энергии радаром относительно большое ($=Magn_2$).

Эта верная, но стилистически неудачная конструкция может быть перифразирована таким образом, что получится ГСС, соответствующая, в частности, переводу:

Радар потребляет относительно большое количество/много энергии.

Этот стилистически более совершенный вариант обеспечивается следующими правилами синонимичных замен:



В результате применения некоторого семантического правила и соответствующего ему синтаксического правила происходят изменения в некоторой части ГСС. Поскольку меняется лексика и синтаксическая структура, вслед за применением правила необходимо проверить, выполняются ли в новой ГСС, полученной в результате преобразования, те требования, которые записаны в словарных статьях русских слов, стоящих в узлах полученной ГСС. Очевидно, что эту проверку достаточно произвести для тех узлов полученной ГСС, в которых произошли какие-то изменения, и для узлов, вновь появившихся. Тут могут представиться две возможности: либо эти требования удовлетворяются и тогда можно переходить к следующему этапу работы: либо эти требования не удовлетворяются, и в этом случае необходимо продолжить работу блока перифразирования.

В принципе, блоку перифразирования можно ставить две задачи: 1) получить из заданной ГСС все возможные правильные ГСС, которые эквивалентны ей по смыслу; 2) применять правила перифразирования до тех пор, пока не будет получена первая ГСС, для которой выполнены все требования, записанные в словарных статьях русских слов, после чего работу этого блока прекратить.

Ранее опробованная на машине система перифразирования предполагала простой перебор правил *) Однако, чтобы получать правильную структуру (т.е. структуру, в которой выполняются все требования, записанные в словарных статьях) возможно более коротким путем, необходимо применять направленный перебор. При таком переборе, по-видимому, правила будут разбиты на некоторые группы, ответственные за устранение тех или иных видов нарушений правильности ГСС (в свою очередь, будут разбиты на группы и виды нарушений правильности структуры). После этого для каждого вида нарушения правильности структуры можно будет задать путь его устранения, указав соответствующую группу правил.

§ 8. Русский глубинный синтаксис.

Данный блок получает на вход русское ГСП с глубинно-синтаксической структурой, приемлемой с точки зрения всех

*) См. статью Арсентьева 1969.

фильтров блока "Русское перифразирование". На выходе выдается поверхностно-синтаксическое представление русской фразы; при этом реально происходит только преобразование ГСС в ПСС.

Результат работы данного блока определен, вообще говоря, неоднозначно: одной глубинно-синтаксической структуре может соответствовать несколько поверхностно-синтаксических структур; некоторые из этих ПСС могут затем быть отвергнуты фильтрами последующих уровней.

Русская поверхностно-синтаксическая структура (также как и английская) представляет собой дерево зависимостей, в узлах которого стоят лексемы с индексами "смысловых" морфологических характеристик *), а ветви помечены символами русских ОНД.

Примеры русских ОНД:

Предикативное: мы ← обращаем

Агентивное: внимание → читателя, перепечатанный → машинисткой

Отпредложное-отсюзное: в → том, что → отставали

Служебное: то, → что; самый ← умелый

Определительное: параллельные ← линии

1-е комплетивное: модулировать → частоту

2-е комплетивное: обращать (внимание) на (факт).

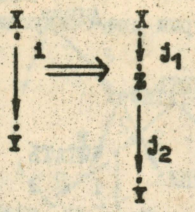
Преобразование ГСС в ПСС осуществляется с помощью формальной грамматики, которая состоит в основном из правил, сводимых к элементарным преобразованиям трех разных типов (i - некоторое глубинно-синтаксическое отношение, j - некоторое поверхностно-синтаксическое отношение):

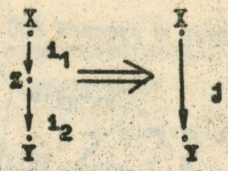
1) X —i→ Y ⟹ X —j→ Y

(замена отношений при сохранении узлов)

---

*) Они отличаются от ГМП отсутствием "синтаксической" морфологической информации, вырабатываемой следующим блоком: § 8 (ср. выше, § 3).

2) 

3) 

(появление нового узла). (исчезновение узла)

Эта грамматика содержит около 100 правил.

Приведем примеры:

1) Глубинно-синтаксическому отношению 1 ("предикат - 1-й актант") соответствуют, в частности, ОНД:

(а) предикативное

(б) определительное и

(в) субстантивно-определительное;

структура АТАКОВАТЬ —1→ ТАНКИ

будет, соответственно, преобразована в

(а) ТАНКИ —предикативное→ АТАКУЮТ

(б) ТАНКОВАЯ —определительное→ АТАКА

(в) АТАКА —субст.-определит.→ ТАНКОВ

(Слова в примерах для ясности даны в том виде, который они получат после морфологического синтеза).

2) Появление нового узла можно показать на примере обработки актантных связей типа 2-4, оформляемых в ПСС - соответственно русской модели управления с помощью предлога:

НАПАДАТЬ —2→ КЕМБРИЯ ⟹ НАПАДАТЬ —1-е объектное→ НА —отпредл.→ КЕМБРИЯ

3) Пример исчезновения узла:

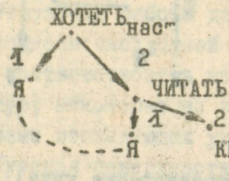
ПРИБЛИЗИТЕЛЬНО —5→ ШЕСТЬ —5→ КНИГА ⟹ ШЕСТЬ —аппроксимативное→ КНИГА (= книг шесть)

В процессе перехода от ГСС к ПСС выполняются две следующие содержательно важные операции:

1. Элиминация тех "повторяющихся", т.е. имеющих тождественные денотаты узлов (и поддеревьев), которые не могут при-

существовать в русской ПСС. Например, рассмотрим две следующие ГСС:

(а)



и (б)

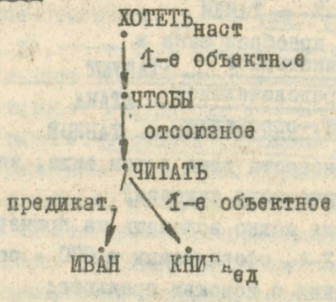


Структурам (а) и (б) соответствуют фразы:

Я хочу читать книгу и

Я хочу, чтобы Иван читал книгу.

В соответствии с нормами русского поверхностного синтаксиса, во втором случае (в отличие от английского языка, где возможен оборот I want John to read the book) поддерево узла ЧИТАТЬ должно быть развернуто в придаточное предложение с подлежащим Иван; соответствующий кусок ПСС будет выглядеть так:

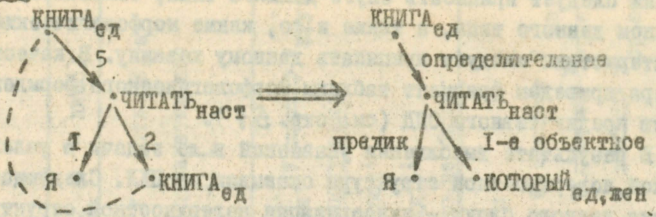


В структуре (а), наоборот, обязательно происходит исчезновение узла – поскольку невозможно *Я хочу, чтобы я читал.

2. Выделение в ПСС фразы тех поддеревьев, которые соответствуют отдельным простым предложениям и приравненным к ним оборотам (причастным и деепричастным, обособленным и т.д.) *) Так, в последнем примере поддерево, висящее на узле

*) Разбиение структуры синтезируемой сложной фразы на более простые компоненты (а иногда – и на отдельные фразы) не воспроизводит механически строение исходной английской фразы, а выполняется с учетом закономерностей только русского языка.

чтобы, отвечает отдельному простому предложению. Аналогично:



где может быть выделено в качестве будущего отдельного предложения поддерево с вершиной ЧИТАТЬ (т.е...., которую я читаю).

Такие поддеревья проходят дальнейшую процедуру синтеза каждое отдельно, а затем результаты соединяются в единую фразу специальным механизмом (см. ниже, стр. 60-61).

§ 9. Русский поверхностный синтаксис.

На вход поступает ПСП русской фразы, а на выходе получается линейно упорядоченная цепочка русских ГМП, задающая фразу *).

Напомним, что в узлах поверхностно-синтаксической структуры стоят ГМП, лишенные некоторых (= "чисто синтаксических") морфологических признаков; данный блок должен превратить их в "полные" ГМП, выработав недостающие им морфологические характеристики. Задача определения этих синтаксически обусловленных морфологических характеристик русских лексем решается с помощью специальных таблиц; для каждого ОНД имеется одна такая таблица.

Эти таблицы имеют следующее строение. Разные столбцы таблицы соответствуют разным типам хозяев, возможных при данном ОНД, а разные строки - разным возможным типам слуг. Под типом здесь имеется в виду набор свойств лексемы (и/или связанных с нею других лексем), влияющих на способ оформления лексем, участвующих в этом ОНД. На пересечении строки и столбца в таблице указывается, какие именно морфологические характе-

*) Образец русской поверхностно-синтаксической структуры дан в Приложении, рис. 4.

ристики следует приписать слуге данного типа, связанному хозяином данного типа, а также и то, какие морфологические характеристики следует приписать данному хозяину. В качестве примера приведем фрагмент таблицы морфологического оформления членов предикативного ОНД (см. стр. 57).

В результате выполнения указанной выше задачи в узлах русской поверхностной структуры оказываются ПХЛ. Следующая функция данного блока – линеаризация поверхностной структуры, т.е. определение порядка расстановки ПХЛ будущей русской фразы. Линеаризация выполняется в два приема: сначала определяется порядок слов в отдельных предложениях, составляющих фразу [1], или в приравненных к ним оборотах, а затем полученные при этом цепочки ПХЛ упорядочиваются друг относительно друга.

Упорядочение слов в пределах простого предложения (деепричастного, причастного оборота и т.п.) разделено на три этапа:

1. Построение исходных групп: определение словопорядка в таких группах, состоящих из хозяина и его непосредственных слуг, в которых словопорядок в основном не зависит от факторов, внешних по отношению к группе. Примеры исходных групп: для этого второго столбца, не только рассматривает, последних два примера.

Порядок слов в исходных группах определяется с помощью так называемых "обойм", задающих относительное расположение слов внутри именных, глагольных, адъективных и адвербиальных сочетаний. В частности, "обойма" словопорядка именной группы имеет (условно) следующий вид:

1	2	3	4	5	6	7	8	9	10	11
и	не ко	толь-	для	всех	этих	трех	на- ших	пер- вых	боль- ших	треу- голь- ников	АВС

[1] Детальное описание правил словопорядка в русском языке дано в работе Мельчук 1965.

Фрагмент морфологического описания членов (ОНД) ИНО

Фрагмент таблицы морфологического оформления членов предикативного ОНД (ОНД I)

Хозяин / Слуга			Личная форма глагола $V_{x,y,w}$ (x – род, y – число, w – лицо)	
			V не при условиях, названных в следующих графах.	V – БЫТЬ, и при V есть НЕ, и $D_2(V)$ – краткое страдательное причастие.
Существительное $S_{x,y,z}$ (x – род, y – число, z – падеж)	Без предлога	Без числительного или с числительным ОДИН, но не МЕНЕЕ/БОЛЕЕ	$S \Longrightarrow S_{им}$, $V \Longrightarrow V_{x,y,w}$ Строка находится...; Я читала. Ты ждешь.	Либо $S \Longrightarrow S_{им}$, $V \Longrightarrow V_{x,y}$: Результаты не были получены; либо $S \Longrightarrow S_{род}$, $V \Longrightarrow V_{ср,ед,3}$; Результатов получено не было.
	С предлогом	С числительным И ≠ ОДИН, где И ≠ 1,	$S \Longrightarrow S_{им}$, и либо $V \Longrightarrow V_{мн,3}$, либо $V \Longrightarrow V_{ср,ед,3}$: Шесть таблиц располагаются/ располагаются справа.	$S \Longrightarrow S_{род}$, и либо $И \Longrightarrow И_{им}$, $V \Longrightarrow V_{мн,3}$ либо $И \Longrightarrow И_{им}$, $V \Longrightarrow V_{ср,ед,3}$ либо $И \Longrightarrow И_{род}$, $V \Longrightarrow V_{ср,ед,3}$ Пять таблиц не были получены; Пяти/пяти таблиц не было получено.

Цифрами обозначены номера позиций. В клетках, для ясности, записаны не символы классов слов (союз, частица, предлог, формула и т.д.), а конкретные представители каждого класса; в позиции 10 помещено существительное **S**, являющееся главным в данной именной группе ("глава группы").

"Обойма" изображает максимально полную именную группу, и тем самым определяет порядок слов для любой неполной группы.

Правила словопорядка в исходных группах являются локальными; расположение слов внутри группы определяется "обоймой" и почти не зависит от контекста (наличие/отсутствие тех или иных классов слов внутри группы, наличие/отсутствие других исходных групп в той же фразе).

2. Построение приведенных групп: линейное упорядочение таких синтаксически связанных исходных групп, взаимное расположение которых определяется правилами, которые условно можно назвать "квази-локальными". Для каждой пары связанных исходных групп их взаимное расположение определено этими правилами однозначно. Если же при одной группе-хозяине имеется несколько групп-слуг, то относительное расположение этих групп-слуг определяется с учетом возможного состава различных приведенных групп. Например, в приведенной группе Алгоритм синтаксического анализа в системе машинного перевода с французского языка на русский, слову алгоритм подчинены две исходные группы - беспредложная (синтаксического анализа) и предложная (в системе...); в соответствии с правилами, первая группа будет поставлена перед второй (а не *Алгоритм в системе машинного перевода с французского языка на русский синтаксического анализа).

В результате построения приведенных групп синтезируемая фраза предстает в виде дерева зависимостей, образованного группами, близкими к традиционным членам предложения: группа личного глагола, которой непосредственно подчинены группа подлежащего, группы дополнений и обстоятельств, а также группа именной или инфинитивной части сказуемого.

3. Взаимное упорядочение приведенных групп должно вы-

полняться с учетом многих крайне разнообразных и сложно взаимодействующих друг с другом факторов, относящихся к фразе в целом. Поэтому здесь трудно сформулировать достаточно жесткие правила типа тех, которые должны использоваться на этапах 1 и 2, в силу чего данный этап будет организован иначе, нежели этапы 1 и 2.

В отличие от чисто локальных правил, действующих на первом этапе, и квази-локальных (то есть учитывающих внутренний контекст образуемой приведенной группы), действующих на втором этапе, правила третьего этапа линеаризации можно назвать "глобальными", так как контекстом, который они должны учитывать при упорядочивании приведенных групп, является вся фраза в целом.

Наиболее сильно действующий фактор на этом этапе - так называемое актуальное членение фразы, то есть распределение темы и ремы, сведения о котором, вырабатываемые при анализе английского текста, содержатся в глубинно-синтаксическом представлении синтезируемой русской фразы (см. выше, § 7). Этому фактору отвечает "абсолютное правило": группа, выделенная как тема, ставится самой левой, группа, выделенная как рема, - самой правой *).

Вслед за этим выполняется расстановка групп дополнений (четырех видов по числу типов глубинно-синтаксической связи), а затем обстоятельственных групп, которых различается приблизительно 20 видов.

Остальные правила, описывающие взаимное расположение приведенных групп, носят скорее характер рекомендаций, корректируемых с помощью специального списка нежелательных ситуаций. В этом списке перечислены типовые случаи неудачных в данном языке словорасположений.

---

*) Ср. пример перевода фраз (1) **The car stopped at a well** и (2) **A car stopped at the well** (рема, определяемая в данных случаях расстановкой артиклей, подчеркнута); русские переводы (1) Машина остановилась у колодца; (2) У колодца остановилась машина.

Приведем несколько примеров нежелательных словорасположений:

– значительно более длинная группа оказывается ближе к группе сказуемого, чем более короткая; например, Специальная комиссия наблюдала за осуществлением регулярного обмена информацией между советскими и американскими специалистами, работающими над словарем, и координацией работ (вместо ... наблюдала за координацией работ и осуществлением...);

– рядом оказывается более трех обстоятельственных групп разного типа; например, Заседание было проведено вечером 6-го июля при участии представителей всех секций с целью обсуждения актуальных вопросов в качестве заключительной встречи участников конференции;

– с одной стороны от группы сказуемого оказывается более двух групп при отсутствии групп с другой стороны от него; например, Снятие омонимии производится на этом этапе с учетом результатов обработки, выполненной ранее, специальной подпрограммой.

Каждой нежелательной ситуации из списка сопоставлено эмпирически подобранное отрицательное число – "штраф". Расстановка слов во фразе в целом оценивается суммой штрафов по всем встречающимся в ней нежелательным ситуациям. Система ищет такого взаимного упорядочения приведенных групп, для которого сумма штрафов за нежелательные расстановки минимальна. Идеальный случай при этом – нулевой штраф. Перебирая варианты, доставляемые множественным синтезом (ср. выше, § 7), система должна стремиться к получению такого идеального варианта.

В том случае, когда добиться нулевого штрафа не удается, система проверяет, нет ли другого варианта ГСП данной фразы (в данном понимании), выданного предыдущим блоком, и если оно есть, пытается проделать процедуру линеаризации над этим ГСП. Сравнив штрафы для разных вариантов, она выберет тот вариант перевода, который дает минимальный штраф. Таким образом, правила словопорядка служат дополнительным фильтром при отборе альтернативных вариантов перевода.

Полученные "линеаризованные" простые предложения и обособ-

ленные обороты объединяются в сложные фразы.

После этого происходит обработка повторяющихся слов – их опущение (эллипсис) и замена местоимениями. Напомним, что все эллипсисы и местоимения, имевшиеся в исходной английской фразе, на этапе глубинно-синтаксического анализа были восстановлены (ср. выше, § 3) [*].

Заключительная операция, выполняемая данным блоком, – расстановка знаков препинания. Она осуществляется в соответствии с нормами русской пунктуации по правилам типа:

I) "Поставить запятые перед и после каждого придаточного предложения и обособленного оборота"; ... I5) "Уничтожить запятые, оказавшиеся рядом с запятой, точкой, вопросительным и восклицательным знаком" и т.п.

§ I0. Русская морфология [**]

На вход данного блока поступает линейно упорядоченная цепочка русских глубинно-морфологических представлений словоформ и знаков препинания. Выходом является цепочка словоформ и знаков препинания, составляющих окончательную фразу.

Переход от ГМП к словоформе, представленной в орфографической записи, выполняется правилами трех типов.

I. Правила перехода от ГМП к цепочке морфем.

Пример:

$$T_{(S)X,Z} \Longrightarrow \{T\} \quad \{YZ\} \quad (Y\text{ –число, }Z\text{ –падеж})$$

По существу это не одно правило, а схема большого числа правил, обеспечивающих получение форм склонения для основ существительного. С помощью подобных правил выполняется, в частности, преобразование:

---

[*] Формальные описания соответствующих разделов грамматики русского языка, на которые будут опираться правила данной группы, имеются: расстановка местоимений исследована Л.Н.Иорданской (Иорданская 1965) и Е.В.Падучевой (Падучева 1971), эллипсис – Н.И.Леонтьевой (Леонтьева 1965), а объединение простых предложений в сложные – Е.В.Падучевой (Падучева 1964).

[**] См. Еськова – Мельчук – Сапогов 1971.

$$\text{ЧИСЛО}_{\text{мн,род}} \Longrightarrow \{\text{ЧИСЛО}\} + \{\text{МН. РОД}\}$$

В результате ГМП, изображенная в левой части, заменяется цепочкой морфем, представленных в правой части правила.

2. Правила перехода от морфем к представляющим их морфам.

Пример: $\{\text{ЧИСЛО}\} \Longrightarrow$ чисел

$\{\text{МН.РОД}\} \Longrightarrow$ ов | муж. П скл. и $m_{-1} \neq$ ...Ш/С' *)

$\Longrightarrow$ ев | либо муж. П и m_{-1} = ...Ш/С' ;
либо жен. Ш скл.

$\Longrightarrow \emptyset$ | либо I скл., либо сред.,
либо муж. П скл. спец.

Справа от вертикальной черты выписаны условия выбора данной правой части при совпадающих левых частях. Для рассматриваемого примера выполняется последнее условие – "сред", что позволяет перейти от исходной цепочки морфем {ЧИСЛО} + {МН.РОД} к цепочке морф

чисел + $\emptyset$.

3. Правила чередований.

Примеры: д ⇒ ж ходить – хожу
ст ⇒ щ пустить – пущу
е ⇒ $\emptyset$ отец – отца

К цепочке морф чисел + $\emptyset$ ни одно из правил чередований не применяется; однако к цепочке чисел + ами применилось бы последнее из приведенных выше правил, что дало бы чисд + ами.

Сведения о необходимости применения правил чередования извлекаются из соответствующего раздела русской словарной статьи.

Результаты работы данного блока должны выдаваться на печать в качестве окончательного результата работы системы АРАП.

IУ. Предполагаемые основные этапы разработки системы АРАП.

Из предыдущего изложения должно быть ясно, что значительная сложность предлагаемой системы делает ее полную реализацию "за один присест" достаточно трудоемкой или даже вообще неосу-

*) Данная запись означает: "предшествующая морфа, т.е. m_{-1}, не оканчивается на шипящую (Ш) или мягкую (С') согласную".

ществимой. В то же время на настоящем этапе исследований можно считать установленным, что обеспечить автоматическое получение высококачественных переводов, используя какие-либо иные средства, существенно более простые, чем предлагаемая система, практически невозможно. Ввиду этого, представляется целесообразным организовать разработку системы таким образом, чтобы уже на промежуточных этапах ее реализации она могла выдавать практически полезные результаты (например, подстрочник к тексту, а также переводы-полуфабрикаты разной степени законченности).

Подчеркнем, что реализация системы АРАП немыслима без мощной организационно-технической базы - достаточного штата лингвистов, математиков, программистов, современных средств вычислительной техники и пр. Стадию кабинетных разработок и первых пробных машинных экспериментов можно считать в основном законченной, и в дальнейшем усилий отдельных энтузиастов будет явно недостаточно для создания настоящей, действующей системы АП. При выполнении этого условия, представляется удобным и практически оправданным подразделить разработку системы АРАП на следующие три этапа:

Первый этап. Создание англо-русского автоматического словаря описанного выше (раздел 1) типа, объемом около 10 тысяч единиц, блока "Английская морфология", а также вспомогательных правил использования словарной информации.

На вход такой системы поступает английский текст, на выходе выдается последовательность русских словарных эквивалентов английских слов. При этом предусматривается возможность разных режимов работы системы: (1) выдавать только русские эквиваленты; (2) выдавать русские эквиваленты плюс их словарные статьи (или отдельные зоны их статей - например, списки лексических функций); (3) выдавать и русские и английские словарные статьи - полностью или частично и т.п.

Практическая полезность автоматического словаря подтверждается опытом эксплуатации таких систем за рубежом -

в ФРГ и Люксембурге *).

Автоматический словарь, составляющий первую очередь системы АРАП, благодаря описанным выше особенностям применяемого в этой системе лингвистического аппарата (лексические функции, модели управления и т.п.) может обеспечить не только выдачу русских эквивалентов, но и отсев тех неверных эквивалентов, которые получаются при пословном переводе несвободных словосочетаний – идиом или комбинаций значения лексической функции с ее ключевым словом.

---

*) В ФРГ автоматический англо-немецкий словарь используется в Бюро переводов Бундесвера (г. Маннгейм). Переводчики, работающие в Бюро, получают тексты, снабженные пристатейным словариком, дающим один или несколько эквивалентов для каждого слова текста, или для отдельных слов, заранее указанных переводчиком; система выдает переводы не только отдельных слов, но и устойчивых (прежде всего терминологических) словосочетаний. Если какой-либо термин отсутствует в автоматическом словаре и не известен переводчику, то переводчик может получить консультацию у специалиста. Выработанный таким образом перевод нового термина сразу же вводится в словарь, так что другие переводчики в дальнейшем будут получать этот перевод без труда. Тем самым, система служит посредником между переводчиками, обеспечивая их эффективное сотрудничество и взаимопомощь. Переводчик может работать также и с текстом, предварительно не прошедшим машинную обработку; тогда переводчик запрашивает систему через телетайп о переводе неизвестных ему слов и словосочетаний. По опубликованным данным применение автоматического словаря, во-первых, позволяет переводчику, не являющемуся специалистом в данной области знания, выполнить терминологически правильные переводы специальных текстов, а во-вторых, сокращает примерно вдвое затраты времени на перевод.

Другая действующая система автоматического словаря – DICAUTOM – осуществляет словарное обеспечение переводчиков текстов по строительной тематике на пяти западно-европейских языках. Выдаваемый этой системой подстрочник позволяет специалисту-строителю, обладающими самыми поверхностными знаниями в языке оригинала (на уровне подготовки, даваемой средней школой), переводить на родной язык тексты по данной тематике.

Поясним сказанное на примере. Пусть имеется английская фраза: **This is particularly important when progress in a development project necessitates changes in one of the routines.**

Русский перевод, который было бы желательно получать от системы АРАП после ее окончательной реализации, таков:

Это становится особенно важным, когда процесс той или иной опытно-конструкторской разработки требует изменений в одной из программ. Набор русских эквивалентов для этой фразы, взятых из словаря, имеет примерно следующий вид:

Этот/это быть/являться особенно важно/важный когда развитие/прогресс в развитие/разработка/создание проект/проектировать/выдаваться/выступать делать необходимым/вызывать необходимость изменение/изменять в один от/из программа.

Построить нормальный перевод фразы по такому набору словарных эквивалентов, по-видимому, нелегко. Это связано не только с обилием вариантов перевода для каждого слова, но и, прежде всего, с тем, что те переводы слов **progress**, **development**, **project** и **necessitates** , которые как раз и нужны в данном контексте, среди предложенных эквивалентов отсутствуют. Автоматический словарь системы АРАП при работе в режиме, учитывающем сложные слова, выдаст правильный перевод для словосочетания **development project** - опытно-конструкторская разработка. При подключении блока "Английская морфология" система припишет словам некоторые морфологические характеристики, которые могут добавочно облегчить понимания фразы:

Этот/это быть/являться$_{наст}$ особенно важно/важный когда развитие/прогресс$_{ед}$ в опытно-конструкторская разработка$_{ед}$ делать$_{наст}$ необходимым/вызывать$_{наст}$ необходимость$_{ед}$ изменение$_{мн}$/изменять$_{наст}$ в один от/из программа$_{мн}$.

Морфологические характеристики несколько проясняют связи между словами, благодаря чему специалисту будет, вероятно, легче перевести данную фразу.

Наконец, при режиме работы, учитывающем, помимо сложных

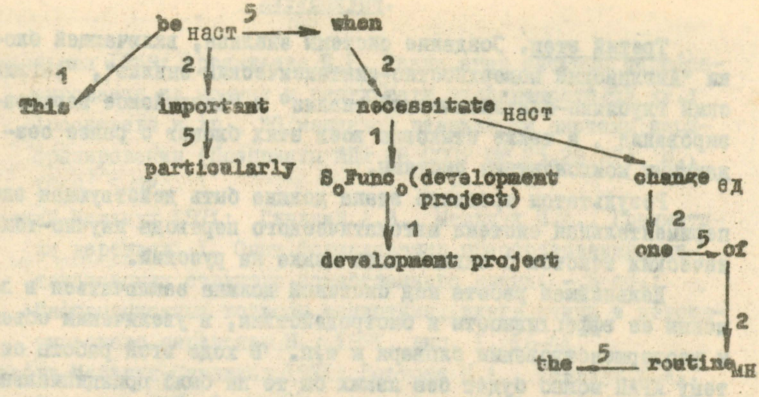
слов, также и сведения, содержащиеся в синтаксических и лексических зонах словарных статей (см. выше, § 1) система выдаст правильный перевод и для слова **progress** . Это слово является значением лексической функции S_0Func_0 от сложного слова **development project** . В модели управления слова **progress** записано, что первое место при нем выражается существительным с предлогом **in** . Эти сведения позволят сначала заменить **progress** символом S_0Func_0 , а затем взять значение этой лексической функции из словарной статьи русского слова РАЗРАБОТКА: процесс. Морфологическое оформление этих слов также изменится: предлог **in** будет отброшен, а слову разработка будет приписана характеристика "род.пад." - в соответствии с указанием, записанным в модели управления русской словарной статьи ПРОЦЕСС (**progress in development project** ⟹ процесс разработки). В результате получится следующий подстрочник, в еще большей степени похожий на правильный перевод рассматриваемой фразы:

Этот/это быть/являться$_{наст}$ особенно важно/важный когда процесс$_{ед}$ опытно-конструкторский разработка$_{ед,род}$ делать$_{наст}$ необходимым/вызывать$_{наст}$ необходимость$_{ед}$ изменение$_{мн}$/изменять$_{3\ ед,наст}$ в один от/из программа$_{мн}$.

Второй этап: (а) Создание всех блоков русского синтеза - выбор русских эквивалентов, русское перифразирование, глубинно-синтаксический, поверхностно-синтаксический и морфологический синтез; (б) создание предварительного варианта системы английского анализа.

Предварительный вариант алгоритма синтаксического анализа, который будет использоваться на этом этапе разработки системы АРАП, должен устанавливать все достоверные синтаксические связи, что покрывает довольно большую долю синтагм, в особенности, в простых и не особенно длинных фразах.

В частности, в рассмотренном примере должна быть в конце концов получена следующая глубинная структура:



На основе этой структуры блоками синтеза может быть построен правильный перевод данной фразы (см.выше).

Использование ГСС и сведений о лексических функциях, содержащихся в словаре, позволяет разрешать омонимию даже в достаточно сложных случаях.

Например, если в данной фразе заменить слово **necessitates**, однозначно определяемое как глагол, словом **demands** (глагол/существительное), то наряду с правильным переводом

... требует изменений в одной из программ,

мог бы быть получен и ошибочный перевод:

... развитие (прогресс) требований опытно-конструкторской разработки изменяет одну из программ.

Однако, поскольку связь между **progress in** и **development project** является "функциональной", т.е. зафиксирована в словаре, а связь между **progress in** и **demands** 'требования' - нет, в соответствии с правилом предпочтения 2 на стр. 37 системой будет выбран первый вариант анализа и, соответственно, перевода.

Тем не менее, разумеется, однозначный и полностью автоматический перевод будет обеспечиваться системой АРАПдалеко не во всех трудных случаях, требующих, в частности, разрешения лексической и некоторых специальных типов синтаксической омонимии.

Третий этап. Создание системы анализа, включающей блоки "Английский поверхностно-синтаксический анализ", "Английский глубинно-синтаксический анализ" и "Английское перифразирование", а также стыковка всех этих блоков с ранее созданными компонентами системы.

Результатом третьего этапа должна быть действующая экспериментальная система автоматического перевода научно-технических текстов с английского языка на русский.

Дальнейшая работа над системой должна заключаться в повышении ее эффективности и быстродействии, в увеличении объема и усовершенствовании словаря и т.п. В ходе этой работы систему АРАП можно будет без каких бы то ни было принципиальных изменений приспособить для перевода текстов любой тематики: это потребует только пополнения словаря соответствующей терминологией.

x

x x

Авторы считают своим приятным долгом выразить свою искреннюю признательность В.М.Григоряну и Т.М.Тер-Микаэляну, прочитавшим рукопись настоящей публикации и сделавшим ряд ценных замечаний.

Литература.

Арсентьева 1969: Арсентьева Н.Г. Реализация системы перифразирования на машине и результаты эксперимента. -В кн.: Арсентьева и др., "О машинной реализации системы перифразирования (Препринты ИПМ АН СССР №№ 25,26,27), М., 1969, 5-82.

Гладкий-Мельчук 1971: Гладкий А.В., Мельчук И.А., Грамматики деревьев. I. Опыт формализации преобразований синтаксических структур естественного языка. - В сб.: Информационные вопросы семиотики, лингвистики и автоматического перевода, М., 1971, вып. I, 16-41.

Еськова-Мельчук-Санников 1971: Еськова Н.А., Мельчук И.А., Санников В.З. Формальная модель русской морфологии. I. Формообразование существительных и прилагательных. (ИРЯ АН СССР. ПГЭПЛ. Предварительные публикации), М., 1970, 71.

Жолковский-Мельчук 1967: Жолковский А.К., Мельчук И.А. О семантическом синтезе. -Проблемы кибернетики, 1967, вып. 19, 177-238.

Иорданская 1964: Иорданская Л.Н. Об одной проблеме автоматического синтеза: употребление личных местоимений 3-го лица в русском языке. -Научно-техническая информация, 1964, № 10, 27-32.

Корельская-Падучева 1971: Корельская Т.Д., Падучева Е.В., О возвратных, личных и указательных местоимениях в русском языке (правила употребления и машинный эксперимент). -Информационные вопросы семиотики, лингвистики и автоматического перевода, М., 1971, 46-66.

Кулагина 1969: Кулагина О.С. Некоторые вопросы преобразования деревьев зависимостей. М., 1969 (Препринт ИПМ АН СССР № 12).

Кулагина-Мельчук-Розенцвейг 1969: Кулагина О.С., Мельчук И.А., Розенцвейг В.Ю. Еще раз к вопросу о реализации автоматического перевода (Ан). -Научно-техническая информация, серия 2, 1969, № 11, 28-32.

Леонтьева 1965: Леонтьева Н.Н. Анализ и синтез русских эллиптических предложений. -Научно-техническая информация, 1965, № 11, 41-46.

Мельчук 1965: Мельчук И.А. Порядок слов при автоматическом синтезе русского текста. -Научно-техническая информация, 1965, № 12, 36-44.

Мельчук 1969: Мельчук И.А. Об одном эксперименте АП с русского языка на французский в рамках Гренобльской системы. -Научно-техническая информация, серия 2, 1969, № 12, 23-29.

Образцы 1970: Образцы словарных статей англо-русского автоматического семантического словаря. (ИРЯ АН СССР. ПГЭПЛ. Предварительные публикации, вып. 9). М., 1970, 58 .

Падучева 1964: Падучева Е.В. Синтез сложных предложений с однозначной синтаксической структурой (при переводе с информационно-логического языка на русский). -Научно-техническая информация, 1964, № 6, 43-49.

Hays 1967: Hays D.J. Introduction to Computational Linguistics. N.Y., 1967, 231 pp.

Kulagina-Mel'čuk 1967: Kulagina O.S., Mel'čuk I.A. Automatic translation of languages: some theoretical aspects and the design of a translation system. -In: Machine translation, ed. A.Booth, 1967, 139-171.

Приложение.

Пример перевода английской фразы: последовательность представлений, получаемых в результате работы отдельных блоков системы АРАП.

В силу многовариантности всех блоков системы на каждом уровне должно получаться несколько альтернативных представлений фразы; однако, чтобы не перегружать пример, мы приводим для каждого уровня только одно представление - именно то, которое требуется для получения правильного результата.

АНГЛИЙСКАЯ ФРАЗА.

We draw reader's attention to the fact that the prime requirement on the modulator is said to be that the phase of the oscillation of the magnetron in the parallel type machines is retarded according to the running list data which is difficult to discover and which specifies what kind of pulse frequency modulates the magnetron.

РУССКИЙ ПЕРЕВОД, КОТОРЫЙ ЖЕЛАТЕЛЬНО ПОЛУЧИТЬ.

Мы обращаем внимание читателя на следующий факт. Утверждается, что основное требование к модулятору таково: отставание фазы колебания магнетрона в машинах параллельного типа должно происходить в соответствии с данными рабочего списка, которые трудно обнаружить и которые определяют, импульс какого типа модулирует частоту магнетрона.

РЕЗУЛЬТАТ АНГЛИЙСКОГО МОРФОЛОГИЧЕСКОГО АНАЛИЗА.

We draw$_{V,3sg,praes}$ reader$_{S,sg,poss}$ attention$_{S,sg}$ to$_{Prep}$ the$_{Art}$ fact$_{S,sg}$ that$_{conj}$ the$_{Art}$ prime requirement$_{S,sg}$ on$_{Prep}$ the$_{Art}$ modulator$_{S,sg}$ be$_{V,3sg,praes}$ say$_{V,pp}$ to$_{Part}$ be$_{V,inf}$ that$_{Conj}$ the$_{Art}$ phase$_{S,sg}$ of$_{Prep}$ the$_{Art}$ oscillation$_{S,sg}$ of$_{Prep}$ the$_{Art}$ magnetron$_{S,sg}$ in$_{Prep}$ the$_{Art}$ parallel$_{A}$ type$_{S,sg}$ machine$_{S,pl}$

be$_{V,3\ sg,praes}$ retard$_{V,pp}$ according to$_{Prep}$ the$_{Art}$ running$_{A}$ list$_{S,sg}$ data$_{S,sg}$ which$_{Conj}$ be$_{V,3\ sg,praes}$ difficult$_{A}$ to$_{Part}$ discover$_{V,inf}$ and$_{Conj}$ which$_{Conj}$ specify$_{V,3\ sg,praes}$ what$_{Conj}$ kind$_{S,sg}$ of$_{Prep}$ pulse$_{S,sg}$ frequency$_{S,sg}$ modulate$_{V,3\ sg,praes}$ the$_{Art}$ magnetron$_{S,sg}$.

БИБЛИОТЕКА
ИН-ТА РУССКОГО ЯЗЫКА
ИН-ТА ЯЗЫКОЗНАНИЯ АН СССР

Ответственный за выпуск Гончар Л.

Корректор Панфилова Т.

Т-03687 от 24.09.71 4,5п.л.+4вкл. 180экз.Зак.№797
Тип.МГПИИЯ

ИРЯ
институт русского языка
им. в. в. виноградова



www.ruslang.ru

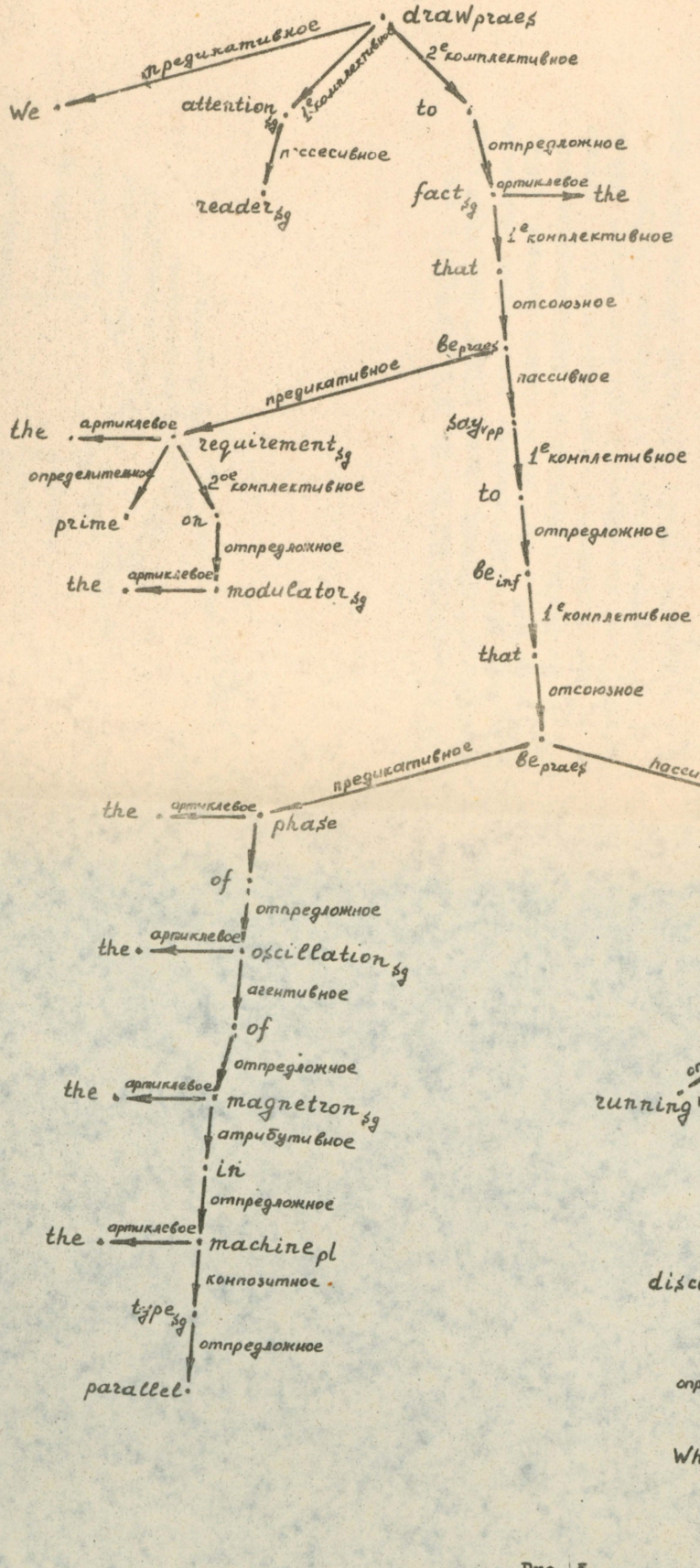


Рис. 1

Замечание 1. При построении ПСС данной фразы группа in the parallel type machines формально может быть отнесена не только к magnetron, но и к oscillation, и к phase. Данную неоднозначность разрешить без понимания сути дела, скорее всего, невозможно. Однако в этом, к счастью, нет и необходимости: ее можно абсолютно точно "воспроизвести" в русском тексте (... фаза колебания магнетрона в машинах параллельного типа...), оставив выбор на долю компетентного читателя. Именно такое решение должно быть предусмотрено в системе.

Замечание 2. Для слова frequency хозяином может быть не только modulates (= pulse frequency modulates , т.е. 'импульс частотно моделирует'), то и of (= of pulse frequency , 'импульсной частоты'). Такую неоднозначность наличными средствами снять невозможно, и она поведет к двум русским переводам: тот, что указан на стр. 71, и ... определяют, импульсная частота какого типа моделирует магнетрон.

АНГЛИЙСКАЯ ГЛУБИННО-СИНТАКСИЧЕСКАЯ СТРУКТУРА

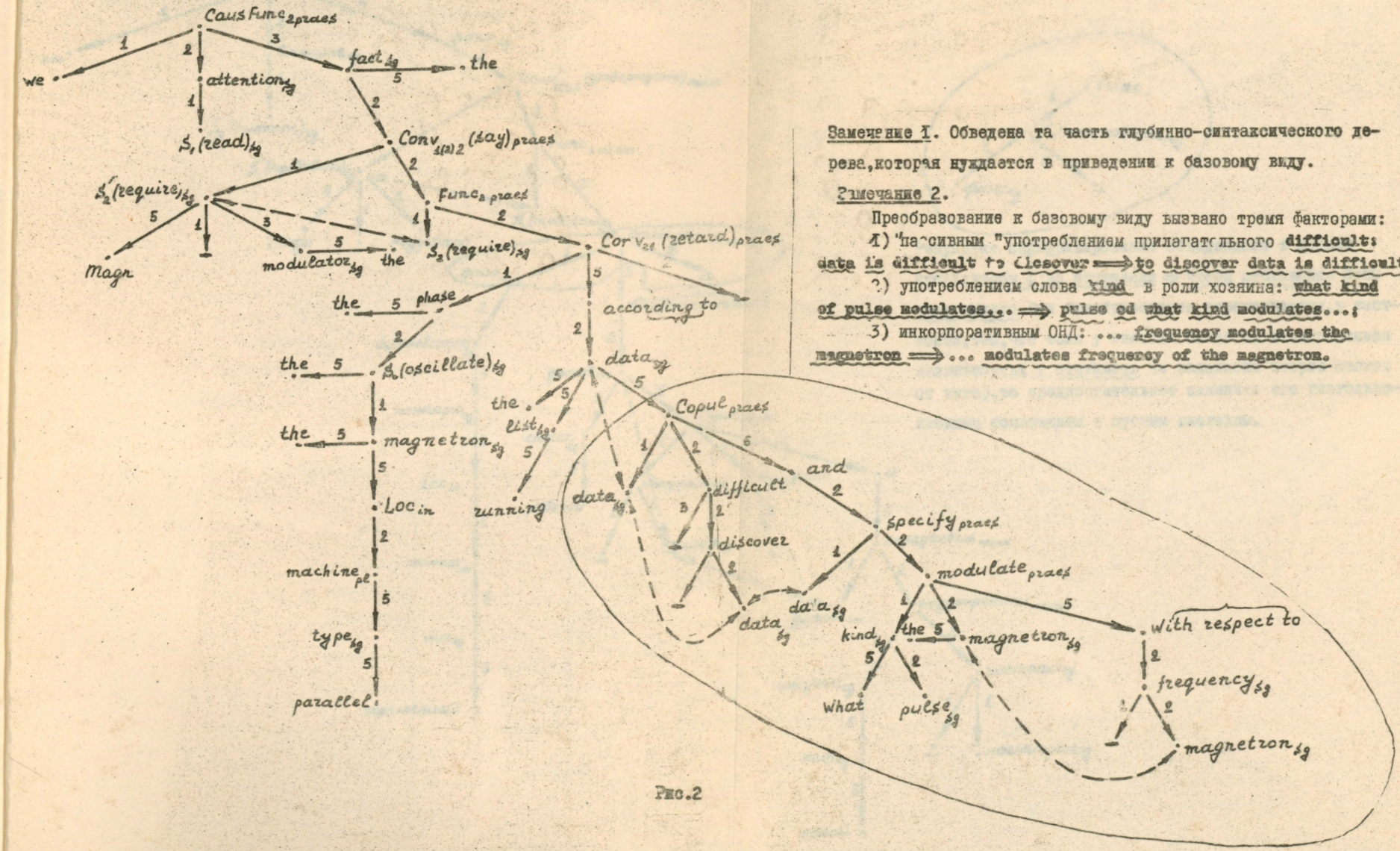


Рис.2

Замечание 1. Обведена та часть глубинно-синтаксического дерева, которая нуждается в приведении к базовому виду.

Замечание 2.

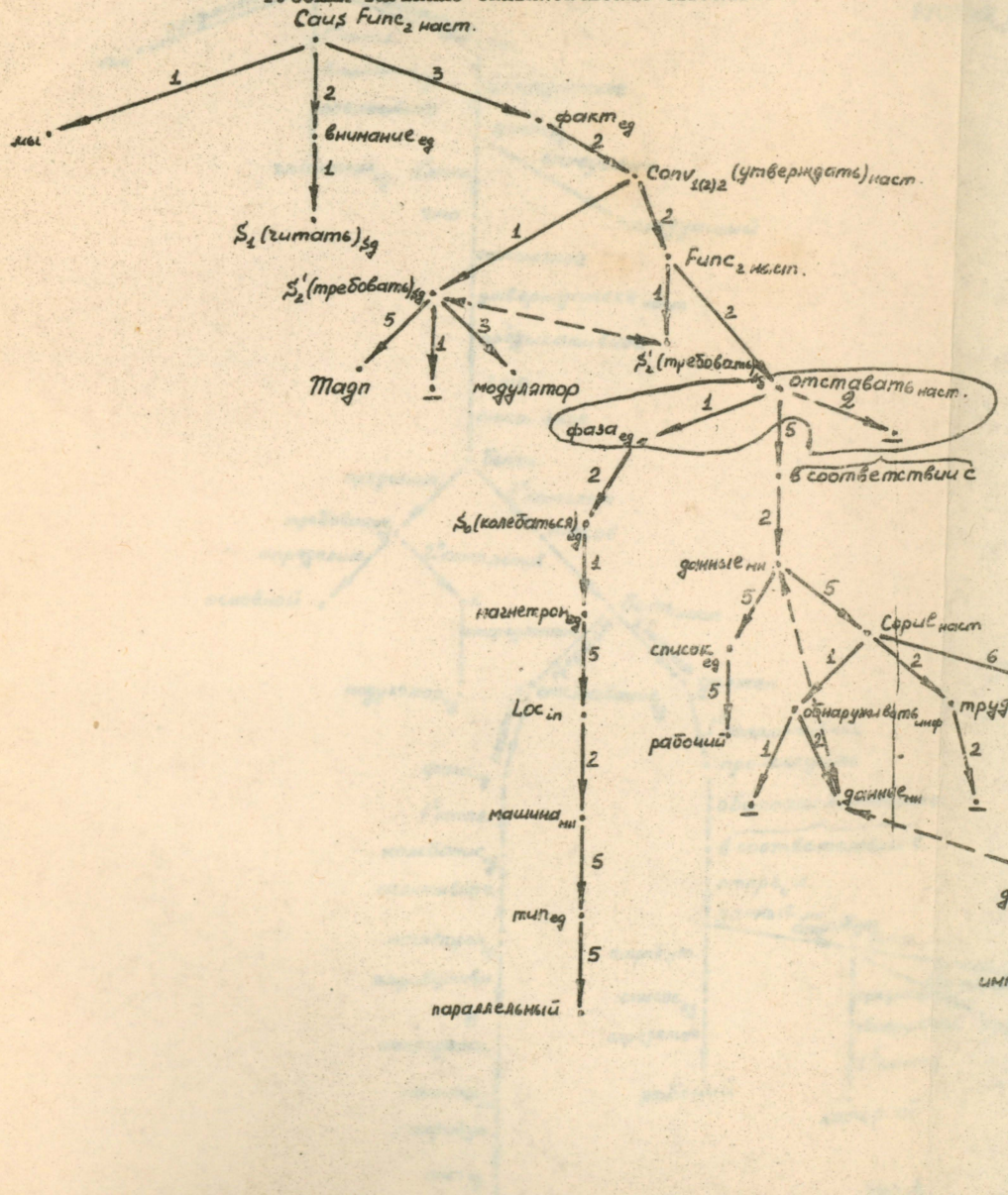
Преобразование к базовому виду вызвано тремя факторами:

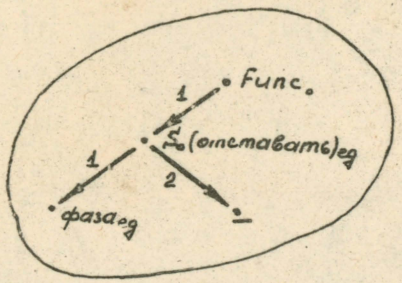
1) пассивным употреблением прилагательного **difficult**: **data is difficult to discover ⟹ to discover data is difficult;**

2) употреблением слова **kind** в роли хозяина: **what kind of pulse modulates... ⟹ pulse od what kind modulates...;**

3) инкорпоративным ОНД: **... frequency modulates the magnetron ⟹ ... modulates frequency of the magnetron.**

РУССКАЯ ГЛУБИННО-СИНТАКСИЧЕСКАЯ СТРУКТУРА





Результат перифразирования обведенного в дереве куска: чтобы фаза отставала ⇒ чтобы отставание фазы происходило. Это преобразование оправдывается, в частности, тем, что если у глагола остается незаполненная валентность (у отставать не заполнено второе место: от чего), то предпочтительнее заменять его глагольно-именным сочетанием с пустым глаголом.

Рис. 3

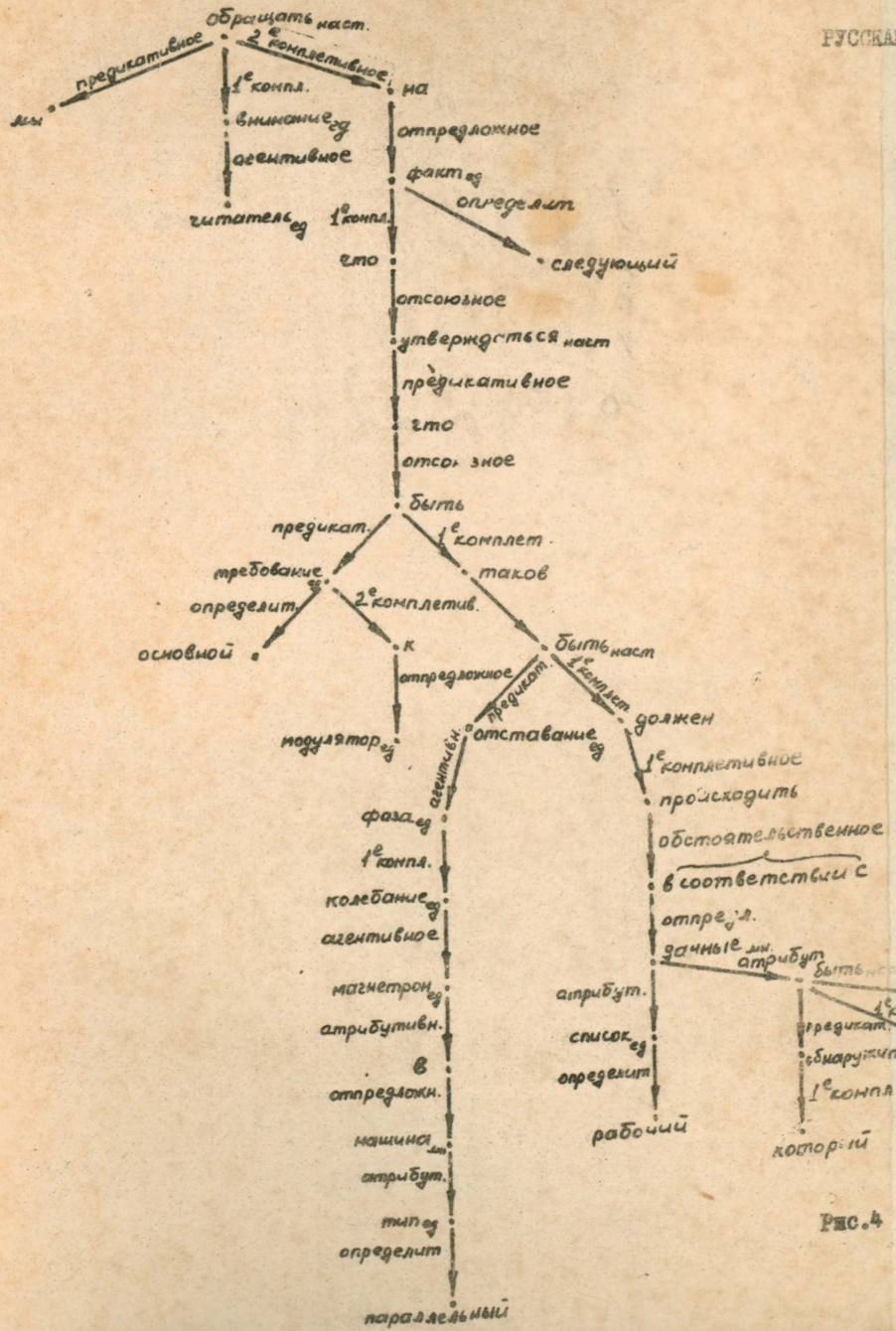
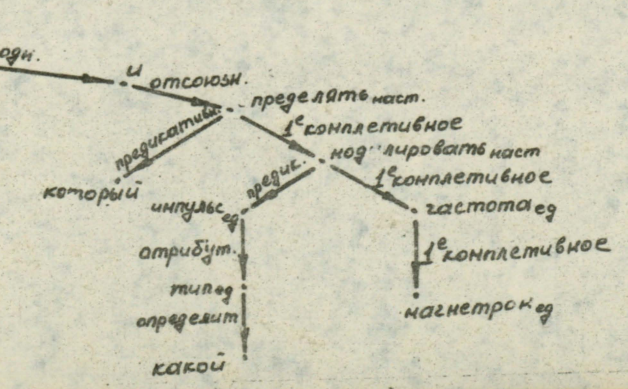


Рис.4



Замечание 1.

При переходе к данной русской ПСС выполняется разбиение ее на три отдельные дерева – для уменьшения громоздкости выходной фразы. Это делается на основании правил, разрешающих отделять и превращать в самостоятельную фразу слишком длинные дополнения – придаточные к существительным типа факт, требование и т.п.; эти же правила добавляют к факт определение следующий, заменяют сослагательное наклонение, зависящее от требование, словом должен (требование, чтобы он ушел ⟹ требование таково: он должен уйти) и т.п.

