

О. С. Кулагина

**О современном
состоянии машинного
перевода**

Рекомендуемая форма библиографической ссылки:
Кулагина О. С. О современном состоянии машинного перевода // Математические вопросы кибернетики. Вып. 3. — М.: Наука, 1991. — С. 5–50. URL: <http://library.keldysh.ru/mvk.asp?id=1991-5>

О СОВРЕМЕННОМ СОСТОЯНИИ МАШИННОГО ПЕРЕВОДА

О. С. КУЛАГИНА

(МОСКВА)

Введение

Настоящая статья представляет собой обзор работ по машинному переводу в мире, характеризующий их состояние к 1988 г. Статья делится на две части. В первой дается общее представление о системах машинного перевода, их типах, строении, организации работы с ними, а также кратко излагается история работ по автоматизации перевода. Во второй части характеризуется состояние работ в разных странах.

Статья рассчитана на читателя, не являющегося специалистом в области автоматизации перевода, в ней не ставилась цель глубокого изложения проблематики.

Данный обзор в значительной мере основывается на материалах Международных конференций по вычислительной лингвистике COLING за десятилетие 1978—1988 гг., конференций Европейского отделения Ассоциации по вычислительной лингвистике и некоторых других конференций, а также на обзорах Дж. Слокума [116] и У. Хатчинса [65]. Сведения о коммерческих системах и вообще о коммерческой стороне дела почерпнуты в основном из этих двух обзоров, а также из сборника [82].

I. МАШИННЫЙ ПЕРЕВОД: ОПРЕДЕЛЕНИЕ, ОЧЕРК ИСТОРИИ

I.1. Об определении машинного перевода

Термином «машинный перевод» (МП) обозначают, с одной стороны, перевод текстов (письменных или устных) с одних естественных языков на другие при помощи ЭВМ, с другой стороны, направление научных исследований, связанных с построением переводческих систем.

В переводе с помощью ЭВМ существует много разных форм организации взаимодействия ЭВМ и человека.

Первоначально предполагалось, что работа переводческих систем должна заменить человека-переводчика, но при этом человек-редактор будет по-прежнему редактировать перевод. Иначе говоря, мыслилась та же организация, что и при «человеческом переводе», т. е. организация с постредактированием.

Другой способ организации работы, который возник несколько позднее, — это перевод с предредактированием, когда редактор исправляет, упрощает или размечает исходный текст, т. е. работа редактора предшествует переводу.

С развитием интерактивного взаимодействия с ЭВМ возникла еще одна разновидность распределения труда между человеком и машиной:

перевод с интерредактированием, когда человек вмешивается в работу переводческой системы на промежуточных стадиях, разрешая трудные случаи. О переводе с интерредактированием правильнее говорить как о частично автоматизированном («машинно-человеческом» — МЧП). Однако непроходимой границы между разными видами организации работ нет, системы с разными видами редактирования имеют много общего, поэтому сейчас нередко термин «машинный перевод» понимается широко, охватывая системы с пост-, пред- и интерредактированием.

В зарубежной литературе употребляются термины и сокращения machine translation (MT), machine-aided или machine-assisted translation (MAT) и сокращение M(A)T, когда хотят охватить и то, и другое.

В настоящий обзор вошли системы всех указанных выше видов.

Близко к системам машинного перевода, но несколько особняком стоят так называемые автоматические словари (иначе — «человеко-машинные» системы). Это системы, где переводит человек, а машина проводит только поиск в словарях. Поскольку в труде человека-переводчика работа со словарями занимает по некоторым оценкам до 40 % времени, то и такая автоматизация может быть полезной, но тут все-таки вся существенная часть работы делается человеком. Автоматические словари не включены в данный обзор систематически, хотя некоторые очень крупные будут упомянуты.

Проиллюстрируем сказанное выше о разных способах распределения труда между человеком и ЭВМ с помощью схемы, на которой прямоугольниками показана часть работы, которую делает ЭВМ, овалами — часть, которую выполняет человек. Ромб в овале показывает, что человек работает с помощью автоматизированных средств редактирования

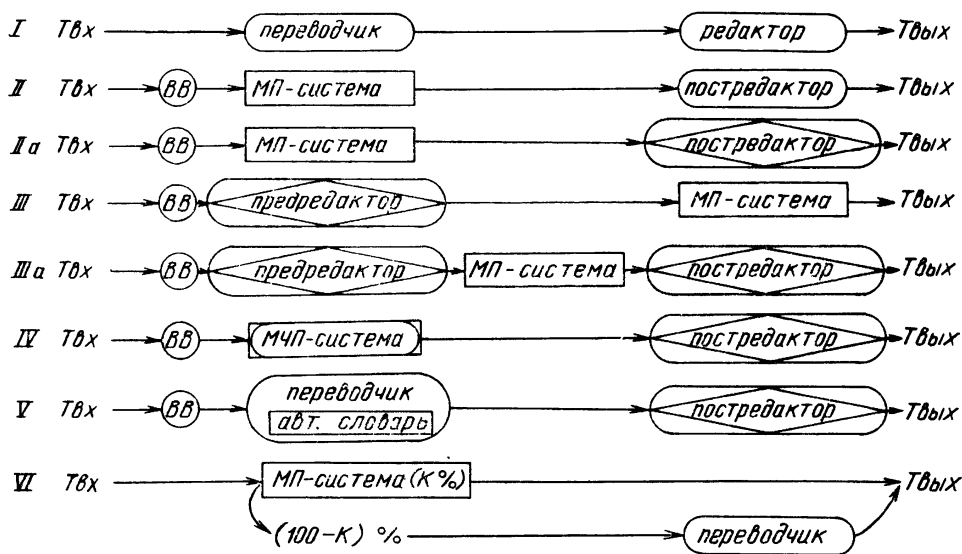


Рис. 1

(text processing). ВВ обозначает выполняемую человеком работу по вводу входного текста ($T_{вх}$) в ЭВМ; $T_{вых}$ — выходной (переводящий) текст.

Строка I рис. 1 — это схема «человеческого» перевода без помощи ЭВМ, строка II — перевода с постредактированием, как он мыслился вначале, IIa — как он организован теперь. Строка III — перевод с предредактором. Строка IV иллюстрирует работу машинно-человеческой системы, а строка V — человеко-машинной. Строка VI относится к организации работы в системе МЕТЕО (см. п. II.12).

Когда используется одна из схем IIa, IIIa, IV, V, разные части работы могут выполняться на разных ЭВМ; часто МП-системы делаются на больших машинах, а для редактирования могут использоваться персональные. Правда, в последнее время стали появляться переводческие системы, рассчитанные на персональные ЭВМ.

1.2. Коротко об истории развития работ по МП

Мы начнем с краткого изложения внешней стороны развития работ по МП.

Машинный перевод насчитывает уже около четырех десятилетий. Мысль о возможности использовать ЭВМ для перевода была высказана еще в 1946 г., сразу после появления первых вычислительных машин. Толчком к активному повсеместному увлечению этой проблематикой послужил известный Джорджтаунский эксперимент (1954 г.). Это был первый в мире публичный эксперимент по переводу, выполненный в Джорджтаунском университете США. Несмотря на его примитивность (словарь в 150 единиц, грамматика из шести правил, перевод нескольких простых предложений), этот эксперимент получил очень широкий резонанс. Пятидесятые и начало шестидесятых годов стали временем возникновения направления и увлечения машинным переводом во многих странах. Начались исследования в Англии, Болгарии, ГДР, Италии, Китае, Франции, ФРГ, Японии и других странах. В том же 1954 году начались работы по МП в СССР.

К концу первого десятилетия работ — к середине шестидесятых годов две системы МП, обе сделанные в США для перевода с русского языка на английский, были представлены для практического использования: одна — система MARK — в Департаменте иностранной техники ВВС США, другая — система GAT, сделанная в Джорджтаунском университете, начала использоваться в Национальной лаборатории атомной энергии в Окридже и в центре Евратома в г. Испра в Италии [55].

Конец периоду первого подъема положило решение печально знаменитой комиссии ALPAC [32], созданной в США, которая, оценив результаты созданных к тому времени систем, пришла к выводу, что, в силу низкого качества даваемых этими системами переводов, МП нерентабелен в условиях США. Правда, в своих выводах комиссия рекомендовала проводить углубленные теоретические исследования для построения в будущем более совершенных систем МП. Однако в целом выводы комиссии были восприняты пессимистически, и это привело к резкому упадку интереса к проблеме МП, к сокращению финансирования, а кое-где и к прекращению работ.

Тем не менее, хотя и в меньшем объеме, исследования по МП продолжались в ряде стран, в том числе и в США. Появлялись новые идеи, подходы (о них подробнее см. в п. 1.3). В пользу возрождения МП действовал ряд объективных факторов. Наиболее существенным из них было активное развитие вычислительной техники. Возможности компьютеров быстро возрастали, и ЭВМ все чаще рассматривались не только как орудие вычислений, но как средство переработки информации в более широком плане. Развивалось создание информационно-поисковых систем, терминологических словарей, терминологических банков данных и т. д. Наконец, появление персональных ЭВМ, изменивших отношение к компьютерам у огромного числа людей, не могло не сказаться и на машинном переводе. Другим сильнодействующим фактором была все возрастающая потребность в переводах в связи с развитием различных видов сотрудничества между разными странами.

Все это привело к тому, что после примерно десятилетия спада в машинном переводе, начиная с середины семидесятых годов, наметился

новый подъем. Появился целый ряд новых практически действующих систем, о которых подробнее будет сказано ниже. Накопился некоторый опыт их эксплуатации; он стал предметом обсуждения на конференции «Практический опыт в машинном переводе», которая прошла в Лондоне в ноябре 1981 года.

Восьмидесятые годы стали временем перехода к широкому практическому использованию переводческих систем, число таких систем стало значительным, сложился рынок коммерческих систем.

Здесь надо отметить, что никому из создателей систем машинного перевода не удалось в полной мере реализовать те мечты, с которыми соответствующие работы начинались, т. е. не удалось создать систем, которые давали бы высококачественный перевод широкого круга текстов. Однако оказалось, что те схемы организации труда, которые изображены на рис. 1 под номерами IIa, IIIa, IV, V позволяют переводчикам работать гораздо быстрее, чем при «классической» схеме под номером I. Это ускорение оценивается в разных переводческих бюро по-разному: в 2, в 3 раза, а при ограничении на входные тексты — и в 5 раз [110, 113].

В конце восьмидесятых годов активное развитие МП продолжается. Увеличивается и число коммерческих систем перевода, и число стран, где ведутся соответствующие работы. Так в восьмидесятых годах начались работы по МП в Испании, Индии, Индонезии.

Нужно сказать, что расстановка сил теперешнего подъема отличается от того, что было в начальный период. Если начало работ по МП было положено в США (первая конференция по МП в 1951 г., Джорджтаунский эксперимент, первые практические системы), то в восьмидесятых годах мировым лидером в МП (как по числу центров, где ведутся исследования, так и по числу специалистов) стала Япония.

В обзоре [65] число специалистов, работающих в области МП в Японии, оценивается в 800—900 человек, что примерно равно числу специалистов по МП во всем остальном мире. Дело в том, что потребность в переводах в Японии сильно отличается от потребности в переводах в США. В первую очередь это объясняется спецификой японского языка, который не имеет распространения вне Японии, в то время как английский в настоящее время распространен в мире настолько, что число научно-технических публикаций, выходящих в разных странах на английском языке, составляет от двух третей до трех четвертей. Япония затрачивает на переводы огромные средства. Недаром японское правительство рассматривает перевод как жизненную необходимость для страны. Вместе с тем чрезвычайно высокий уровень развития вычислительной техники в Японии и хорошо известный проект машин пятого поколения способствуют невычислительным приложениям ЭВМ и, в частности, МП.

1.3. Краткая характеристика эволюции подходов к проблеме МП

В п. 1.2 мы изложили внешний ход событий в области МП, охарактеризуем теперь кратко внутреннюю эволюцию области.

Создание переводческой системы требует решения обширного круга проблем из разных областей знания. Требуется: а) определить для каких языков строится система; б) определить круг переводимых текстов; в) установить требования к качеству результата; г) выбрать способ распределения труда между человеком и ЭВМ (см. выше); д) выбрать общую стратегию, т. е. тип переводческой системы; е) разработать разные виды обеспечения системы. Решения перечисленных проблем не независимы.

При выборе языков, учитываемых в системе, возникают следующие вопросы: число языков, выбор входного (входных) и выходного (выходных) языков.

Первоначально эти вопросы обычно решались следующим образом: два языка, причем перевод ведется с некоторого чужого языка на язык страны, где делается система (т. е. на «свой» язык). Это перевод с целью получения информации. Со временем все больше внимания стал привлекать перевод со своего языка на чужой с целью распространения информации. Стали создаваться системы, в которых участвуют более двух языков. Тут встречаются разные комбинации: один входной язык и несколько выходных, реже — несколько входных при одном выходном; возникли также системы, в которых для группы языков предполагается перевод с любого на любой.

С самого начала работ по автоматизации перевода круг переводимых текстов мыслился как ограниченный: речь почти всегда шла о переводе текстов научно-технического и делового характера (объединяемых иногда под названием «деловой прозы»), редко — о переводе публицистики. Иными словами, вопрос о переводе художественной литературы, а тем более — поэзии, не ставился. Причем до недавнего времени речь шла только о переводе письменных текстов.

Наиболее частыми случаями были: при переводе с целью получения информации (т. е. с чужого языка на свой) — перевод научно-технических статей (иногда только заглавий, аннотаций, рефератов) и патентов, а при переводе с своего языка на чужие — перевод разного рода технической документации (инструкций и т. п.), сопровождающей экспортируемую в другие страны промышленную продукцию.

Указанные ограничения на вид текстов всегда сопровождаются ограничениями и на тематику, т. е. на определенную предметную область (например, речь идет о переводе статей по металлургии или по медицине, руководств по вычислительной технике и т. п.).

Несколько особняком стоят системы, рассчитанные только на очень специфический вид входных сообщений, т. е. с сильными ограничениями не только на тематику, но и на строение входов. Такова система МЕТЕО (см. п. II.12), переводящая только тексты метеосводок.

В последнее время возникло новое направление в МП, о котором будет подробнее говориться ниже. Это новое направление — перевод диалога, включая устный диалог. Здесь возникает целый ряд новых обстоятельств. Во-первых, пока в этих системах предполагается только разговор по определенной очень узкой теме: например, резервирование места в гостинице, выяснение условий участия в конференции, т. е. выяснение ее сроков, тематики, формы представления материалов и т. п. Такого рода темы при переводе письменных текстов не рассматривались. Во-вторых, при устном диалоге система перевода должна сочетаться с системой анализа и/или синтеза устной речи. О других особенностях перевода диалога см. в п. II.2.

Что касается требований, предъявляемых к качеству результата, то они оказываются связанными с целью осуществления перевода. Конечно, вообще говоря, всегда лучше получать перевод хорошего качества. Но, как показал опыт использования уже первых практических систем (см. [55]), в тех случаях, когда перевод делается с целью получения информации, даже необработанный постредактором результат работы системы может оказаться полезным пользователю, знакомому с предметной областью, совершенно не знающему языка, с которого переводят, и не имеющему средств достаточно быстро и дешево получить перевод другим способом. В описанной ситуации плохой перевод может быть достаточным для вынесения суждения о том, содержит ли данный текст какую-то полезную для потребителя информацию. Русский читатель

легко может представить себе подобную ситуацию, вообразив, что перед ним два текста: один на японском или китайском в иероглифах, а другой — пусть на сколь угодно корявом русском языке.

Ситуация перевода на чужой язык, т. е. с целью распространения информации, обычно требует более высокого качества перевода, но тут дело облегчается большими ограничениями на входные тексты (см. например, в п. II.15 о системе SYSTRAN в фирме Xerox).

О возможных способах распределения труда между человеком и ЭВМ уже шла речь в п. I.1, мы будем говорить об этом также в ч. II, разбирая различные системы.

Рассмотрим теперь, что подразумевается под типом переводческой системы.

Когда перевод выполняется человеком, то, по-видимому, он происходит по следующей очень общей схеме. Переводя очередной фрагмент текста (фразу, абзац и т. д.), человек уясняет себе его смысл, а затем передает этот смысл на выходном языке, соблюдая при этом не только смысловую, но и определенную синтаксическую близость к оригиналу (иначе это не перевод, а пересказ). При этом он пользуется как лингвистическими знаниями, т. е. знаниями об обоих языках, участвующих в переводе, так и экстралингвистическими знаниями, т. е. знаниями о предметной области, об общих закономерностях окружающего мира, о законах коммуникации и т. п.

Как ни заманчиво было бы воспроизвести такую схему с наибольшей полнотой в переводческой системе, при существующем уровне знаний и возможностей ЭВМ это неосуществимо. Естественно, что при построении переводческих систем идут по пути построения тех или иных приближений к этой общей схеме.

Заметим, что почти во всех ныне существующих системах перевод ведется отдельными фразами, поэтому ниже будет идти речь об анализе или синтезе отдельной фразы.

Первые системы МП, получившие впоследствии названия «систем первого поколения» или «прямых (direct) систем» — П-систем, производили переход от языка к языку на уровне, очень близком к поверхностному и весьма далеком от смысла.

Перевод в П-системе происходит в общих чертах следующим образом. Сначала переводимая фраза преобразуется в ее морфологическое представление. В нем каждая текстовая единица: словоформа, знак препинания или оборот (словосочетание, которое должно обрабатываться не пословно, а как единое целое), — заменяется совокупностью признаков. Это осуществляется при помощи словаря основ, в котором каждой основе соответствует набор признаков, и словаря оборотов, хранящего сведения об оборотах, а также при помощи морфологического анализа (МАН), в процессе которого по окончанию (части словоформы, оставшейся после выделения в ней основы) устанавливаются сведения о форме слова (морфологические характеристики).

Затем морфологическое представление входной фразы перерабатывается в морфологическое представление выходной фразы, т. е. путем анализа формы слова и его линейного контекста устанавливаются сведения о форме и месте переводящего слова, иначе говоря происходят одновременно и анализ, и преобразование входной фразы, ориентированные на выходной язык.

Заключительный этап — это морфологический синтез (МСинт), где на основе данных, записанных в морфологическом представлении выходной фразы, из словаря выходного языка берутся требуемые слова, им придаются требуемые формы и они расставляются в указанном порядке. Все трудности перевода при таком подходе сосредоточены на этапе перехода от морфологического представления входа к морфологическому

представлению выхода, а в целом описанная схема близка к получению подстрочника.

Следует отметить, что в П-системах правила обычно формулировались в виде инструкции для выполнения определенных действий, т. е. грамматика и алгоритм работы с нею были слиты воедино. Когда в 1957 г. В. Ингве провозгласил «отделение грамматики от механизма» [138], это было существенным шагом вперед в развитии переводческих систем.

Опыт работы с системами прямого типа лег в основу создания более сложных систем.

В так называемых Т-системах (от transfer — трансфер*) после морфологического анализа следует синтаксический анализ (САН), в ходе которого морфологическое представление входной фразы преобразуется в ее синтаксическое представление (обычно в виде синтаксического дерева). Затем следует этап преобразования, который в этом случае складывается из преобразования дерева входной фразы в дерево выходной и из смены лексики. Далее на этапе синтаксического синтеза (ССинт) полученное синтаксическое представление перерабатывается в морфологическое, т. е. вырабатываются сведения о порядке и форме переводящих слов. МСинт в Т-системах такой же, как в П-системах.

Шестидесятые годы были годами увлечения синтаксическим анализом, чему немало способствовало появление широко известных формальных моделей и языков Н. Хомского.

Влияние идей Хомского на МП трудно оценить однозначно. С одной стороны, они внесли совершенно непривычные в лингвистике точность и строгость, с другой стороны, как показало время, они не были адекватным аппаратом для описания естественных языков.

Грамматики Хомского были ориентированы на получение синтаксической структуры предложения в виде дерева составляющих: иерархической совокупности вложенных групп слов, где самая крупная составляющая совпадает с фразой в целом, а самые мелкие — со словоформами. Вложение составляющих, устанавливаемое по определенным правилам, определяло синтаксические отношения между ними.

Синтаксический анализ, опирающийся на наиболее распространенные — контекстно-свободные грамматики Хомского, шел обычно снизу вверх, т. е. постепенно собирались все более крупные составляющие. При этом, в сущности, проведение анализа было организовано как проверка синтаксической правильности входного предложения с позиций данной грамматики, а дерево синтаксического анализа получалось только в том случае, если такая правильность имела место. Однако понимание синтаксического анализа как распознавания правильности оказалось неудовлетворительным в качестве основы для автоматического синтаксического анализа по ряду причин. Во-первых, даже для носителей языка граница, отделяющая синтаксически правильные предложения от синтаксически неправильных, является нечеткой, например, из-за нечеткости границ синтаксиса и семантики. Во-вторых, из-за наличия в естественных языках множества слов со специфическими особенностями оказалось слишком трудным создавать описание грамматики такое, чтобы в рамках этой грамматики оказались правильными хотя бы все те предложения, правильность которых для носителей языка не вызывает сомнений. О неадекватности грамматики Хомского для естественных языков см., например, [137].

Более широкими возможностями, чем были у распознавателей для контекстно-свободных грамматик, обладали так называемые расширен-

*) Мы берем «кальку», чтобы не путать с нетерминологическим употреблением переводов данного слова «преобразование», «переход», «перевод».

ные сети переходов Вудса (ATN — augmented transition network [137]), но и они имели свои недостатки с точки зрения анализа естественных языков (об этом подробнее см. [13]).

Другой вид синтаксических деревьев, получивший широкое распространение, — это деревья зависимостей. В этих деревьях все узлы соответствуют текстовым единицам (см. выше), а дуги — синтаксическим связям между ними.

Для деревьев зависимостей известны рассчитанные на них различные способы проведения синтаксического анализа, из которых наиболее удачным представляется так называемый фильтровый метод (коротко о нем сказано в п. II.11, а подробно он разобран в [13]).

Со временем основное внимание при конструировании систем перевода постепенно переместилось с этапа синтаксического анализа на этап преобразования от языка к языку — этап трансфер. Сюда перешло преодоление многих трудностей перевода: преодоление различий в синтаксических структурах, выбор переводных эквивалентов и др. От этого этапа Т-системы и получили свое название, иногда их называют также системами второго поколения.

Третий этап переводческих систем — это И-системы (от *interlingua* — интерлингвовые), их называют еще системами третьего поколения. В идеале И-системы — это системы, в которых конечным результатом анализа является некоторое представление, единое для входной и выходной фразы, свободное от специфических особенностей входного и выходного языков. Получение интерлингвового представления, вообще говоря, требует не только знаний о семантике и прагматике соответствующих языков, но также привлечения экстралингвистических знаний об устройстве мира и о предметной области, к которой относится текст. Реально же на сегодняшний день не существует даже достаточно полных описаний семантики естественных языков, которые охватывали бы и семантику текста (не только предложения), не говоря уже о прагматике, а тем более — об экстралингвистических знаниях (хотя бы в объеме, в котором ими владеет средний специалист). Поэтому в тех переводческих системах, которые строились и строятся как И-системы, после синтаксического анализа следует семантический анализ, дающий семантическое представление, которое считается единым (с точностью до лексики) для входного и выходного предложения, хотя оно и не содержит полной информации.

Семантическое представление имеет обычно вид сети, не являющейся деревом. На основе семантического представления с лексикой уже из выходного языка в процессе семантического синтеза формируется синтаксическое представление выходного предложения, после чего следуют синтаксический и морфологический синтез, такие же как в Т-системах. Ввиду сказанного, представляется, что было бы правильнее называть подобные системы семантическими, а не И-системами.

Следует отметить, что исторически смена типов систем развивалась последовательно. После того, как был сделан ряд П-систем, были предприняты попытки построить сразу И-системы. Наиболее последовательно такая попытка была проведена в жизнь в Гренобльском университете во Франции (см. п. II.6). Однако выяснилось, что на переводе отрицательно сказывался тот факт, что получаемое переходное представление не было настоящим интерлингвовым, в частности в нем не отражались сведения о коммуникативной организации текста, выражаемые, например, порядком слов, и другие сведения, для извлечения и учета которых еще не было выработано средств. Иными словами, эта попытка оказалась преждевременной, она не была подкреплена надлежащим развитием семантики, прагматики и когнитологии. В силу этого на долгое время вплоть до сегодняшнего дня наиболее распространенным типом переводческих систем стали Т-системы, хотя, как будет видно из характеристики от-

дельных систем, приводимой ниже, некоторые разработчики мыслят свои системы как интерлингвовые.

Сказанное выше о типах систем можно для наглядности проиллюстрировать схемой, приведенной на рис. 2.

Как уже было сказано, наибольшее распространение в настоящее время имеют Т-системы. Для перехода от входного языка к выходному в них используются различные представления переводимого сообщения. Иными словами, уровень перехода от языка к языку (уровень преобразования, трансфер) может располагаться по-разному, он может быть поверхностно-синтаксическим, глубинно-синтаксическим, синтактико-семантическим и др. (на рис. 2 эти разновидности не отражены). Вопрос о глубине уровня перехода от языка к языку решается в зависимости от

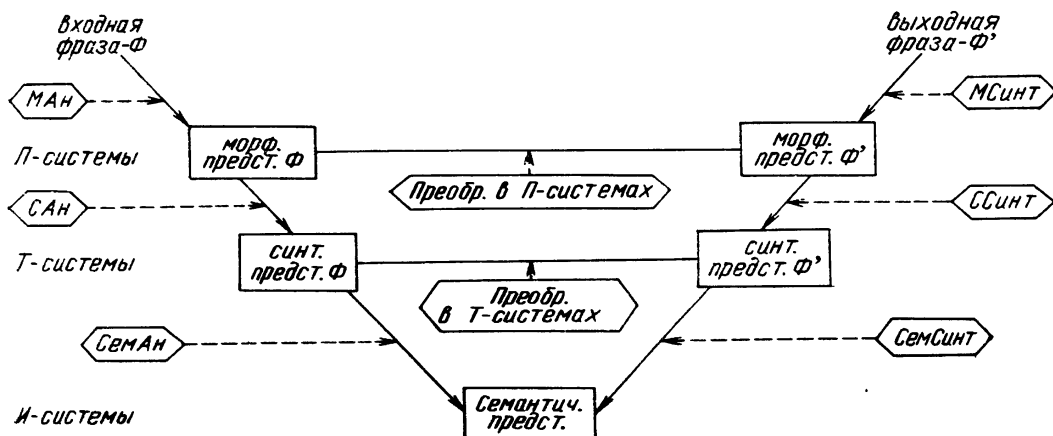


Рис. 2

того, каковы входной и выходной языки. При близко родственных языках переход может совершаться на уровне, более близком к поверхности (примеры будут приведены ниже). Наоборот, чем существеннее различия между языками, тем глубже должен располагаться уровень перехода.

Собственно построение переводческой системы состоит в разработке разных видов ее обеспечения. Практически все переводческие системы имеют три вида обеспечения: 1) лингвистическое, 2) математико-алгоритмическое, 3) программное. Иногда они дополняются еще информационным и логическим видами обеспечения.

Поясним, что относится к каждому из видов обеспечения.

Лингвистическое обеспечение включает в себя словари, таблицы суффиксов и окончаний, нужные для морфологического анализа и синтеза, и грамматические правила для остальных этапов работы.

Математико-алгоритмическое обеспечение включает в себя формализмы для представления лингвистических и перерабатываемых данных и алгоритмы, определяющие работу правил.

Программное обеспечение содержит программы, реализующие работу самой переводческой системы, а также обслуживающие программы (для поддержки словаря, внесения изменений в правила и т. п.).

Информационное обеспечение — это совокупность экстралингвистических данных, например, данных о конкретной предметной области.

Логическое обеспечение — это совокупность логических правил вывода, с помощью которых экстралингвистические данные применяются для уточнения представлений обрабатываемого текста.

Выше мы уже упомянули о том, что в ранних системах МП разные виды обеспечения нередко объединялись: грамматические правила описывались в виде инструкций, предписывающих проведение определенных действий с перерабатываемыми данными. Однако уже с шестидесятых

годов разные виды обеспечения стали разрабатываться согласованно, но самостоятельно.

Разработка перечисленных видов обеспечения невозможна без проведения многочисленных экспериментов по переводу. Поэтому сначала создается экспериментальная система, содержащая исходный основной набор правил, который пополняется, исправляется, усовершенствуется путем проведения экспериментальных переводов. Такая стадия как правило требует нескольких лет работы.

Если цели работы — исследовательские, то система может так и остаться экспериментальной. Если цель — получение практически действующей системы, а тем более, коммерческой, то отработанная в эксперименте система становится прототипом для тиражирования, которое может происходить с изменением программного обеспечения или приспособлением словарей к потребностям пользователя.

Практически действующая система обычно включается в одну из приведенных на рис. 1 схем организации работы, но иногда необработанный результат, выданный системой, поступает непосредственно пользователю — специалисту в соответствующей предметной области, которому достаточно получить общее представление о содержании текста.

При обзоре переводческие системы можно классифицировать по стадиям развития (экспериментальная, прототип, практическая, коммерческая), но эта характеристика меняется со временем. Их можно классифицировать по способам решения тех или иных проблем (моделям грамматик, типам структур и т. п.), но это требует более детального описания систем, чем позволяют рамки данного обзора, рассчитанного на неспециалистов. Поэтому был выбран «географический» принцип построения обзора.

II. СОСТОЯНИЕ РАБОТ ПО МП В РАЗНЫХ СТРАНАХ

II.1. О порядке изложения

Ниже описывается состояние работ по МП в разных странах Азии, Европы и Америки. Правда выдержать принцип «по странам» удастся не вполне. Во-первых, существуют международные проекты, где в разработке участвуют специалисты многих стран. Во-вторых, коммерческие системы делаются в одних странах, а опыт их эксплуатации накапливается в других. Крупные международные проекты вынесены в отдельные пункты (II.3, II.5, II.14). Наиболее распространенные коммерческие системы собраны в п. II.15 и II.16.

II.2. Работы по МП в Японии

Как уже говорилось, Япония в настоящее время является лидером в области МП. В Японии широко ведутся как работы по построению практически действующих систем, так и исследовательские разработки. Последние происходят на фоне широкого фронта работ по вычислительной, или компьютерной лингвистике, о размахе которых свидетельствует, например, тот факт, что в Международной конференции по вычислительной лингвистике КОЛИНГ 80, которая прошла в Токио в 1980 г., приняли участие специалисты, представлявшие 85 организаций из 26 городов Японии. Работы ведутся в таких высших учебных заведениях, как университеты Киото, Токио, префектуры Осака, Киуши, Кубе, Оита, Васеда, Фукуока, Технологический институт Токио и др. По оценке У. Хатчинса [65] из общего числа японских специалистов по МП, составляющего 800—900 человек, примерно 60 % работают в различных фирмах. Например, лаборатории по МП имеются в таких корпорациях,

как Fujitsu, Hitachi, Toshiba, Nippon electric company, IBM Japan, Sharp, Sanyo, Nippon Telephone Telegraph, Oki computer corporation и др. При этом работы ведутся не только в крупных фирмах, но и в мелких.

Как правило японские специалисты делают системы для перевода в паре языков: японский — английский, в ту или иную, или в обе стороны. Гораздо реже японский выступает в сочетании с другими языками, например, с немецким (см. ниже).

Крупнейшая из японских разработок в области МП — это японский национальный проект (Mu — проект), финансируемый специальным координационным фондом для развития науки и техники Агентства науки и техники (Agency of Science and Technology), рассчитанный на 1982—1990 гг.

Этот проект имеет целью перевод с японского языка на английский и с английского на японский резюме статей по науке и технике. Основная исследовательская часть в нем делалась специалистами из университета Киото (группа под руководством М. Nagaо). Кроме того, участвуют три другие организации. Electrotechnical Laboratories (ETL) отвечает за ввод и вывод текстов, морфологический анализ и синтез, частично за словари. Japan Information Center for Science and Technology (JICST) также отвечает за словари, в частности за терминологию. Research Information Processing System under Agency of Engineering Technology разрабатывает человеко-машинный интерфейс, средства поддержки словарей и грамматик.

Первая часть проекта выполнялась в 1982—1986 годах. Были созданы системы-прототипы, после чего основная часть работы перешла к JICST, а специалисты из университета Киото начали новые исследования, о которых говорится ниже.

Обе системы перевода (англо-японского и японо-английского) Mu-проекта относятся к T-системам с переходом на уровне синтаксических структур.

Анализ японского входа содержит следующие этапы: морфологический анализ, установление границ простых предложений, установление сферы действия сочинительных союзов, анализ именных групп, анализ простых предложений, анализ вставных относительных придаточных, обработка сведений о времени и виде, анализ семантических связей между простыми предложениями, обработка контекста (в частности интерпретация частиц), преобразование структуры составляющих в структуру зависимостей в терминах глубинных падежей (синтактико-семантических ролей аргументов предикатов). Используется 33 глубинных падежа для японского и 32 — для английского.

Этап преобразования подразделяется на три части: подготовительная часть, когда полученное представление японского предложения несколько стандартизуется и приближается к английскому; собственно преобразование (т. е. переход от представления японского предложения к представлению английского и смена лексики); постпреобразование (переделка в полученном представлении английского предложения вставных придаточных, соотношений между простыми предложениями и др.).

На этапе синтаксического синтеза полученное дерево просматривается сверху вниз. Здесь может возникнуть необходимость в некоторых дополнительных преобразованиях: часто требуется преобразование существительного в прилагательное или глагол в инфинитиве; если опущен агенс, то предложение преобразуется в пассив; если опущены и агенс, и объект, то номинализируется глагол и добавляется связка и т. д.

Многие из преобразований вызваны тем, что между английским и японским есть существенное различие, состоящее в том, что японский — это так называемый VE-language (т. е. язык, где обычно подлежащее — участник ситуации, с которым нечто происходит), тогда как англий-

ский — это DO-language (т. е. язык, где обычно подлежащее предложения — это главный деятель ситуации). Например, одна и та же ситуация будет обрисована в английском предложением «землетрясение разрушило дом», а в японском: «дом был разрушен землетрясением».

Перевод завершается морфологическим синтезом.

При англо-японском переводе анализ английского предложения делится на морфологический анализ, пред-анализ (разбиение предложения на фрагменты), основной анализ (синтактико-семантический анализ фрагментов) и пост-анализ (установление связей между фрагментами). Много внимания уделяется различным правилам снятия омонимии как по линейному контексту (снизу вверх), так и по частичным результатам анализа (сверху вниз). Используется много эвристических правил, упорядоченных по убыванию «силы».

Дальнейшие этапы перевода обычные: преобразование, состоящее из лексического и структурного преобразований, синтаксический и морфологический синтез. При этом структурное преобразование во многом определяется свойствами лексических единиц. Например, для лексических единиц с предикатными свойствами в словаре определен не только переводной эквивалент, но и даны соотношения между валентностями, которые определяют необходимые переподчинения при переходе от дерева английской структуры к дереву японской. Система специальных правил устанавливает видо-временные характеристики, которые в японском сильно отличаются от английских.

Словари систем к моменту окончания первой части проекта содержали по 80 тысяч единиц. При этом в них включены не только *ОСНОВЫ* слов и те или иные наборы признаков, но часто и специфические для отдельных лексических единиц правила работы.

Для записи правил грамматики специалистами из университета Киото был разработан специальный язык GRADE. Правила имеют вид преобразований «дерево — дерево» с гибкой системой записи условий, результирующих преобразований и способами влиять на последовательность работы правил. Интерпретатор правил реализован на ЛИСП на машине FACOM 380.

С 1986 г. началась разработка второй части Му-проекта, в результате которой система-прототип должна быть переработана в действующую. Эта часть проекта рассчитана до 1990 г.

Му-проект описан в [92—95, 122] и других публикациях.

Как уже было сказано, коллективы, работающие в области МП, имеются в ряде университетов Японии.

Группа из университета Киото до работы над Му-проектом разработала узко специальную переводческую систему TITRAN, предназначенную для перевода заглавий научно-технических статей в информационных базах данных. Первоначально был сделан перевод с английского на японский, затем версии для перевода с японского на английский и французский; совместно с университетом Саарбрюкена предполагают приспособить ее для японо-немецкого.

Эта система переводит только именные группы, поскольку заглавия на 98 % состоят из именных групп. Глаголы в английском тексте допускаются только в виде форм на -ing и -ed и в сочетании to + V(inf). Простая именная группа — это существительное с определениями. Замена каждой простой именной группы на ее главу дает скелетную схему предложения. В 10 тысячах заглавий было выявлено 18 различных схем. Простая группа переводится пословно, если она не содержится целиком в словаре словосочетаний. Перестановки происходят на уровне скелетной схемы, т. е. для целых групп. Система TITRAN для широкого класса заглавий обеспечивает точность 80 %, а для заглавий из области математики и физики — 93 % [91].

После окончания первой части Ми-проекта специалисты университета Киото начали работу в совершенно новом направлении: интерактивный перевод диалога в реальном времени.

Как отметили М. Нагао и Дж. Цудзии в [124], перевод диалога имеет ряд существенных отличий от перевода письменных текстов. Сама постановка задачи тут иная, поэтому системы для перевода диалога нельзя строить как развитие систем, предназначенных для перевода текстов.

В настоящее время рассматривается задача перевода диалога по узкой заранее определенной тематике: например, резервирование места в гостинице или регистрация для участия в конференции. В отличие от перевода текстов в данной ситуации не обязательна структурная (синтаксическая) близость к оригиналу, т. е. здесь можно давать не перевод, а довольно свободный пересказ, важно только правильно передать ту информацию, которая существенна для понимания высказываний. В [124] приводится такой пример. На вопрос «В каком районе Токио Вам нужна гостиница?» клиент может ответить: «Я хотел бы остановиться около А» или «Возле А, пожалуйста» или «Где-нибудь недалеко от А» и т. д. (где А — название места в городе). Все эти ответы можно перевести одинаково, важно только передать понятие близости к А.

Известно, что диалог ведется обычно более короткими и простыми фразами, чем пишутся тексты, хотя вместе с тем в диалоге чаще встречаются эллиптические конструкции.

Еще одно существенное отличие состоит в том, что при переводе диалога переводческая система может обращаться к участникам диалога с вопросами для уточнения извлеченной ею информации. Но при этом она должна работать так, чтобы не слишком замедлять разговором с нею основной диалог.

Пока работы по переводу диалога в университете Киото находятся в исследовательской стадии.

В университете префектуры Осака не менее десяти лет ведутся работы по МП. В 1980 г. Ф. Нисида и др. доложили на COLING 80 о системе англо-японского перевода, которая тогда находилась в стадии начального экспериментального опробования. Перевод предполагается через структуру глубинных падежей. В процессе САп сначала обрабатываются выражения с отрицанием и идиоматические обороты, затем строится дерево глубинных падежей, для построения которого используются модели управления английских глаголов из словаря Хорнби [63]. Полученное дерево преобразуется с учетом запретов, свойственных японскому языку (например, нельзя строить предложение, в котором неживое — субъект, а живое — объект), а также отдельных преобразований, приписанных конкретным глаголам [98].

В 1982 г. на COLING 82 тот же коллектив рассказал о системе японо-английского перевода, основанной на таком же подходе.

Несколько позднее эта группа занялась разработкой пакета программ для переводческих систем: Library Module-Aided Program Synthesizing, который к 1986 г. включал 30 базовых модулей и 24 прикладных [99].

В университете Токио делается система для перевода в паре языков японский-корейский в обе стороны (система NARA). Анализ ведется с использованием GPSG (Generalised Phrase Structure Grammar), трансфер сравнительно прост, так как языки в смысле синтаксиса достаточно сходные. К 1986 г. была сделана система-прототип [47].

Имеются другие разработки для перевода в паре японский-корейский, о них см. ниже.

В университете Киуши также ведутся работы по англо-японскому переводу, в частности, по проблеме выбора переводных эквивалентов

([139]). В университете Кобе ведутся работы по анализу персидского. Имеются исследовательские группы МП также в ряде других университетов.

Как уже отмечалось, бóльшая часть японских специалистов по МП работает в различных фирмах.

В фирме Fujitsu работы по МП ведутся с 1978 г. В 1980 г. сообщалось о первых экспериментах по японо-английскому переводу [126]. В [111] система японо-английского перевода названа ATLAS/I.

Однако в 1984 г. под именем ATLAS/I на рынок была выпущена система англо-японского перевода, которая стала первой коммерческой системой, сделанной в Японии. Она является системой прямого типа, где в процессе анализа для каждой английской текстовой единицы выясняется ее синтаксическая роль и на этой основе устанавливается синтаксическая роль соответствующей японской единицы; если надо, вставляются служебные слова. Система реализуется на машинах FACOM M340 (или более мощных моделях FACOM) плюс миникомпьютер S 3000 для постредактирования. Она особенно эффективна, когда имеется большой текст, из которого надо быстро извлечь информацию. К 1986 г. она была установлена в 70 организациях, из которых половина — университеты и институты (переводящие научно-технические отчеты и др.), а остальное — фирмы (переводящие руководства и технические документы и т. п.) [80].

Система ATLAS/II для японо-английского перевода вышла на японский рынок в июне 1985 г. Она является системой типа трансфер. На этапе синтактико-семантического анализа используется описание синтаксиса в виде КС-грамматики, а также некоторая «модель мира» (понятия и отношения). Представление уровня трансфер имеет вид семантической сети. Оно ориентировано на превращение ATLAS/II в многоязычную систему. Словари содержат по 50 тыс. единиц общей лексики и 250 тыс. научно-технических терминов. На машине FACOM 380 скорость перевода — 60 тыс. слов в час. Через три месяца после выхода ATLAS/II на рынок она была установлена в 30 организациях [80].

В [118] сообщается мнение о внедрении переводческой системы одного из покупателей системы ATLAS/II N. Saeki (представителя автомобильной компании Mazda). Он считает, что внедрение требует доработки, которая начинается с экспериментов по переводу текстов разных типов, чтобы выбрать такие, для перевода которых система наиболее эффективна. Затем начинается период приспособления и доводки системы, в первую очередь расширения словарей и анализирующей грамматики. Это может потребовать 1,5—2 года. При этом важно, чтобы программное обеспечение системы соответствовало тому, которое имеется у потребителя, и чтобы он мог проводить расширение словарей, не обращаясь к создателям системы.

На EXPO'85 демонстрировались также экспериментальные версии ATLAS/II для перевода с японского на немецкий и французский и с английского на японский, немецкий, французский [85].

Анализ японского из системы ATLAS/II включен в систему-прототип, сделанную для японо-немецкого перевода заголовков научных статей совместно фирмой Fujitsu и Штутгартским университетом ФРГ [108].

Большая группа специалистов по МП работает в фирме Hitachi Ltd. Начало работ тоже относится к 1978 г. Работы ведутся для японо-английского и англо-японского перевода.

Система японо-английского перевода выпущена на рынок в 1987 г. Она носит название NICATS/JE (Hitachi Computer-Aided Translation System). Эта система имеет словари в 250 000 единиц, предназначена для перевода научной и технической документации, руководств и т. п.

Реализована на языке ПЛ/1 на машинах серии Hitachi M, переводит со скоростью 60 тыс. слов в час. Процесс анализа делится на морфологический анализ, синтаксический анализ, дающий дерево зависимостей, семантический анализ, вырабатывающий глубинные падежи и снимающий неоднозначность при помощи семантических характеристик, преобразование, синтаксический синтез и морфологический синтез. Система предусматривает предредактирование (расстановку скобок при структурных неоднозначностях) и постредактирование (в частности, выбор переводных эквивалентов) в интерактивном режиме.

Система для перевода с английского на японский — HICATS/EJ (ранее называвшаяся ATHENE — Automatic Translation of Hitachi from English into Nihongo with Editor Support) на рынок пока не вышла, хотя система-прототип была реализована уже к 1982 г. Разработчики называют свою модель эвристической: HPM (Heuristic Parsing Model). Анализируемое предложение расчленяется на составляющие, среди которых различаются составляющие-слова (word element), составляющие-группы (phrasal element) и составляющие-простые предложения (clausal element). Они связаны между собой отношениями вложения и модификации. Используются модели управления словаря Хорнби [63], устанавливаются синтаксические роли составляющих. Эта информация служит для перестановки составляющих и вставки частиц на этапе преобразования. Данная система тоже реализована на языке ПЛ/1.

В 1988 г. фирма сообщила о завершении японо-корейской системы, сделанной совместно с Корейским институтом экономики и техники (Korean Institute of Economics and Technology) [65].

В исследовательском центре формы Toshiba автоматизируется перевод в паре языков английский — японский в обе стороны. Система англо-японского перевода начала продаваться в 1987 г. под названием AS-TRANSAC (ранее — TAURUS: Toshiba Automatic Translation System Reinforced by Semantics). Она работает на миникомпьютерах AS 3000 Toshiba со скоростью 5—7 тыс. слов в час; программы на языке Си.

Это двуязычная трансфер система, рассчитанная на пакетный режим. Процесс перевода делится на поиск в словаре, морфологический анализ, синтаксический анализ, который может дать омонимичную структуру (о нем см. ниже), семантический анализ совмещенный с преобразованием (при котором снимается синтаксическая и лексическая омонимия, в частности путем широкого использования специфических правил сочетаемости, записанных в словарных статьях), синтаксический и морфологический синтез.

Японо-английская версия в 1988 г. находилась еще в стадии **работки**.

Возможности этих двух систем были продемонстрированы при переводе диалога в реальном времени в сеансе связи через спутник между лабораторией фирмы Тосиба в Японии и посетителями V Международной выставки телекоммуникации Telecom-87, которая проходила в октябре 1987 г. в Женеве. Оператор в Швейцарии набирал на терминале сообщение на английском языке, оно передавалось в Японию, где переводилось системой на японский язык. Японский оператор видел на своем экране японский перевод и печатал ответ на японском. Этот ответ тут же переводился системой и его английский перевод передавался в Женеву.

При синтаксическом анализе используется некоторая модификация расширенной сети переходов Вудса (ATN — Augmented Transition Network, [137]). Расширение состоит в возможности «заглядывать вперед» при анализе. Если из некоторого узла могут выходить несколько дуг, которым можно приписать условия о необходимости наличия во фразе определенных элементов, то, проверяя наличие необходимых элементов

в еще не проанализированной части фразы, можно часто отбросить некоторые дуги, не проводя полного анализа. Этот способ позволяет существенно сократить число возникающих возможностей и тем самым уменьшить время анализа. В [101], где описывается данная система, приводится такой пример: 1) I heard the lecture, 2) I heard the lecture was interesting. Группа «the lecture» может быть подлежащим придаточного предложения, только если правее нее имеется глагол. Здесь надо заметить, что все приводимые примеры неявно опираются на то, что поступающие для анализа предложения будут сравнительно простыми и короткими. Видимо, для диалога такое предположение оправдано.

Некоторые другие способы приписывания условий позволяют отбрасывать целые группы дуг. В целом, как утверждают авторы, им удалось вместо затрат времени, оцениваемых $K_1 C^n$ (где n — длина анализируемого предложения, K_1 и C — константы), получить время порядка $K_2 n$. Правда, авторы понимают, что при переводе текстов экономия может быть меньшей.

Система PIVOT фирмы NEC (Nippon Electric Company) отличается от основной массы переводческих систем, тем, что она по своим установкам интерлингвая. Сделана для англо-японского перевода в обе стороны, имеет словари с 40 тыс. единиц для японского и 53 тыс. единиц для английского, покрывающих 20 предметных областей. Процесс перевода делится на морфологический, синтактико-семантический и концептуальный анализ, синтаксический и морфологический синтез. Анализ дает интерлингвовое представление в виде сети, узлы которых — семантические примитивы, а дуги помечены символами глубинных падежей, причем эта сеть отражает не только семантическую информацию, но и прагматическую (деление на тему и рему). В системе используется база экстралингвистических данных и логический вывод. Она реализована на языке Си на микрокомпьютере ACOS System 4, продается с 1986 г. Снабжена средствами поддержки словаря, а также средствами для пред- и постредактирования [61, 65]. В NEC ведутся также работы по автоматизации устного диалога.

В г. Осака в ATR Interpreting Telephony Research Laboratories проводятся работы по переводу устного диалога, ведущегося по телефону. Исследовательский проект рассчитан на 15 лет. Система перевода устного диалога включает в себя подсистемы распознавания речи, собственно перевода и синтеза речи. Известно, что системы распознавания речи работают с ограниченной точностью, поэтому для системы перевода возникают дополнительные трудности, связанные с уточнением входного сообщения. Особенно трудно распознаются короткие служебные слова, среди которых много близких по звучанию, а длинные значащие слова распознаются лучше. Основываясь на этом, разработчики системы (см. [73]) предложили способ уточнения распознавания путем построения правильного предложения из поступивших значащих слов и сравнения его служебных слов с теми гипотезами, которые дало распознающее устройство. Тематика диалога установлена заранее: конференция (сроки, темы, условия регистрации, резервирование гостиницы и т. п.). Пока работы находятся в начальной стадии.

В NTT (Nippon Telegraph and Telephone) с 1981 г. ведутся работы над системой LUTE (Language Understander Translator Editor) предназначенной для англо-японского перевода в обе стороны. Эта система явно нестандартная: в ней предполагается перевод через понимание текста, точнее через построение представления для смысла текста, причем предполагается использование не только лингвистических, но и экстралингвистических знаний (общих — common-sense и экспертных).

В словарях слова сопровождаются разветвленной информацией как синтаксической (включающей фреймы в терминах глубинных падежей),

так и семантической (включающей семантические категории и их отношения).

Структуры отдельных предложений строятся в виде фреймов, из которых конструируется фрейм всего текста. Используется богатый набор семантических отношений: 35 отношений для имен (часть — целое, материал — предмет, обладатель — предмет и т. д.), 33 падежных отношения (отношения между предикатом и субъектом, объектом и т. д.), 3 отношения для вставных предложений (отношение между модифицируемой именной группой и глаголом модифицирующего предложения и др.), 27 отношений между предложениями (условие, причина, результат и т. д.).

Основной формализм для представления как перерабатываемых данных, так и разнообразных используемых лингвистических и экстралингвистических данных — фреймы, точнее сеть фреймов.

К 1986 г. была реализована экспериментальная система для японо-английского перевода со словарями по 3000 единиц. Программы были написаны на языке Zeta Lisp для машины Symbolics Lisp machine [102].

Эти работы имеют в основном исследовательские цели, в частности, выяснение соотношений между лингвистическими и экстралингвистическими данными.

Лаборатории корпорации IBM с 1982 г. разрабатывают систему для перевода с английского на японский руководств по машинам IBM, рассчитанную на внутреннее пользование. Цель разработки — получить высококачественный перевод в ограниченной предметной области с учетом того, что переводимые руководства пишутся довольно стандартным языком. В 1986 г. была сделана система-прототип [125]. Система типа трансфер, причем анализ может давать несколько разных структур с оценкой степени правдоподобия. При снятии неоднозначности на этапе преобразования используется набор семантических признаков (24 семантических класса), отражающий специфику текстов.

В Oki computer corporation делается система PENSEE типа трансфер. Ее японо-английская версия вышла на рынок в 1986 г., реализована на языке Си на микрокомпьютере. Англо-японская версия разрабатывается. Результат нуждается в существенном постредактировании, а японский вход — в предредактировании.

Система фирмы Sanuo для японо-английского перевода продается с апреля 1987 г. Система типа трансфер, имеет словарь 55 тысяч единиц, предусматривает пред- и постредактирование.

О следующих трех системах У. Хатчинс [65] пишет, что их выход на рынок запланирован на 1988 г., а Г. Сомерс в [118] говорит о них, как уже продающихся в 1987 г.

Система MELTRAN фирмы Mitsubishi для японо-английского перевода может работать в пакетном и интерактивном режиме. Реализуется на языке Пролог.

Система RMT фирмы Ricoh для англо-японского перевода реализуется на языке Си. Там же разрабатывается японо-английская версия.

Система фирмы Sharp для англо-японского перевода реализуется на языке Си.

Все эти три системы типа трансфер, работают на микрокомпьютерах, требуют существенного постредактирования и иногда предредактирования [65].

Имеется еще ряд разработок в разных фирмах, делающих переводческие системы для внутреннего пользования.

Помимо работ по построению переводческих систем в Японии ведутся также серьезные работы по построению больших словарей, ориентированных на системы автоматической обработки текста. Мы здесь укажем две работы этого плана.

В EDR (Electric Dictionary Research Institute) имеется большой EDR-проект, который финансируется вместе с государством 8 крупными фирмами по электронике (Fujitsu, Hitachi, Mitsubishi, Matsushita, NEC, Oki, Sharp, Toshiba).

Проект предусматривает создание словарей и математического обеспечения для их поддержки. Словари предполагаются двух типов. Словарь слов должен содержать 200 тыс. слов общей лексики и 100 тыс. терминов из области информационных процессов для японского и английского языков. Слова должны быть снабжены определениями и грамматическими признаками. Словарь понятий должен содержать сведения о соотношениях понятий: зависимостях, синонимии, гипернимии, гипонимии и другие, принятые в тезаурусах. Словари предполагается использовать как для переводческих, так и для информационно-поисковых систем. Большое внимание уделяется проблемам единообразия, согласованности, точности описаний слов и понятий [65].

Другая работа со словарями, на которой стоит остановиться,— это работа, которая делается в университете Киото уже упоминавшимся коллективом под руководством М. Нагао.

В качестве исходного материала берется LDOCE (Longman Dictionary of Contemporary English)—словарь, рассчитанный на использование человеком и имеющийся на машинных носителях. В нем слова снабжены толкованиями. В [96] рассказано об автоматическом построении тезауруса на базе LDOCE. Для этого анализируются толкования слов и из них извлекаются ключевые глаголы (из толкований глаголов) или ключевые и функциональные имена (из толкований имен). Составлен список из 41 функционального слова, объединяющихся в группы: 'вид' (kind, sort etc), часть (part, side, top etc), принадлежность к множеству (member, set, class, group etc) и другие. Толкование с ключевым именем имеет вид

(det) (adj*) KEY NOUN (adj phrase*),

с ключевым и функциональным

(det) (adj*) FUNCTION NOUN of KEY NOUN (adj phrase*).

Знак * указывает возможность повторения.

Выделение указанных элементов толкования позволяет устанавливать различные отношения между словами словаря. Например, собирать в группы глаголы с общим ключевым глаголом. Так, kick = to hit with the foot и knee = to hit with the knee имеют общий ключевой глагол hit.

Работы на базе LDOCE с целью получения сведений, нужных для автоматической обработки текстов, ведутся также в США [135].

II.3. Проект ODA

Недавно японское правительство начало финансировать новый крупный проект ODA (Overseas Development Agency), начатый в 1987 г. и рассчитанный на шесть лет исследований и разработок. Кроме Японии, в нем участвуют Китай, Тайланд, Малайзия, Индонезия.

Проект имеет весьма амбициозные цели: построить многоязычную интерлингвовую систему, которая должна переводить с предредактированием со скоростью 5000 слов в час и точностью 80—90 %. В проекте *участвуют до 200 исследователей из 10 центров. Предполагаются языки: японский, китайский, малайский, индонезийский, тайский*, причем в первую очередь будут делаться версии для пар языков, где один из языков — японский.

Этот проект имеет цель способствовать развитию технических и культурных связей Японии с другими азиатскими странами, в частности, в области компьютерной техники [65, 118].

II.4. Работы по МП в других странах Азии

Китай. Работы по МП в Китае начались в ряде организаций свыше тридцати лет назад. Так, еще в 1956 году тема «машинный перевод» вошла в первый национальный проект развития науки Китайской Народной республики. В 1957 году начались работы по русско-китайскому и англо-китайскому переводу, а в 1959 г. были продемонстрированы первые эксперименты по переводу с русского на китайский.

Однако затем наступил период застоя в работах, охвативший десятилетие с середины шестидесятых до середины семидесятых годов. Среди прочих причин он был вызван недостаточным развитием лингвистических исследований для преодоления различия между китайским и индоевропейскими языками, трудностями китайской иероглифики и недостаточной мощностью компьютерной базы.

В 1975 г. началось возрождение интереса к МП. Работы по МП вошли в очередной пятилетний план. За прошедшее с тех пор время началась разработка шести систем для англо-китайского перевода, двух — для русско-китайского, одной — для французско-китайского и одной для перевода с китайского на несколько языков. Две последние разработки велись в контакте с группой GETA (Франция), о которой будет рассказано ниже.

Главной организацией, занимающейся МП в Китае, является Лаборатория инженерной лингвистики Китайской корпорации по математическому обеспечению (Language Engineering Laboratory — LEL of China Software Technique Corporation — CSTC).

LEL ответственна за МП-проекты, включенные в седьмой пятилетний план (1986—1990), среди которых основными являются проекты систем англо-китайского и японо-китайского перевода. Она ответственна также за сотрудничество с Японией в проекте ODA. Работы по МП ведутся также в Институте лингвистических исследований Китайской Академии общественных наук, Институте научной и технической информации Китая, Институте вычислительной техники, Институте математического обеспечения, университетах Нанкина, Шанхая и др., Политехнических институтах Харбина, Хуангана и др.

В 1987 г. LEL продемонстрировала и выпустила на продажу систему TRANSTAR, которая используется в переводческих бюро. Это система типа трансфер для англо-китайского перевода, имеет словарь, содержащий 40 тысяч слов общей лексики и 40 тысяч терминов. Реализована на языке КОБОЛ на машине IBM AT, снабжена средствами постредактирования. Терминологическая часть словаря ориентирована на тексты из области военных наук, электроники, химии, экономики.

Ряд других систем находятся в стадии экспериментального опробования.

Комиссии по МП, способствующие развитию исследований, имеются в Обществе научной и технической информации Китая, Китайском обществе информационных процессов. В восьмидесятых годах в Китае прошел ряд национальных и международных конференций по вопросам автоматизации перевода. Многие китайские специалисты были посланы для изучения работ по переводу в другие страны: во Францию (Гренобльский университет), в США (университет Техаса) [54, 65].

Малайзия. Работы по МП в Малайзии в University Sains Malaysia начались в 1976 г. как усилия одиночек и стали активнее развиваться после установления контактов с французской группой GETA в 1979 г.

Первоначальной целью было построение системы-прототипа для перевода с английского языка на малайский учебников по химии для средней школы. Система базировалась на математическом обеспечении ARIANE, разработанном в Гренобле (см. п. II.6). К 1986 г. был по-

строен экспериментальный вариант, на основе которого в 1986—1988 гг. должен был быть построен прототип для практической системы. В частности, словари должны были быть доведены до 10 тысяч единиц. Предполагается также расширить систему для перевода текстов по компьютерным наукам [121, 65].

Южная Корея. Выше уже упоминалось о некоторых совместных разработках Японии и Южной Кореи. В настоящее время работы по МП в Южной Корее наиболее активно ведутся в SERI/KAIST (Systems Engineering Research Institute of the Korean Advanced Institute of Science and Technology), в Высшей школе KAIST, в университетах Сеула и других городов.

Многие из них делаются для перевода в паре корейский-японский в ту или иную сторону. Кроме упомянутых в п. II.2, имеются следующие. В институте SERI/KAIST сделали систему для корейско-японского перевода KANT/I, работающую на микрокомпьютере, и начали делать систему японско-корейского перевода совместно с японской фирмой Fujitsu. Высшая школа KAIST провела фундаментальные исследования по анализу корейского и начала совместно с NEC (Япония) работу над многоязычной системой, включающей корейско-японский и корейско-английский перевод. В Сеульском университете делается система корейско-английского перевода, а в сотрудничестве с фирмой IBM — система англо-корейского перевода. SERI, кроме упомянутых работ, делает системы англо-корейского и французско-корейского перевода в сотрудничестве с группой GETA (Франция), ориентированные на математическое обеспечение ARIANE.

Разрабатываемые в Южной Корее системы, как правило, относятся к типу трансфер-систем. Наиболее продвинута разработка, которая ведется совместно с японской фирмой Fujitsu. Остальные системы в 1988 г. были далеки от перехода к практической реализации [65].

Таиланд. В Таиланде с 1981 г. начались работы по созданию системы перевода с английского на тайский, в которых участвуют университеты Бангкока и другие, и которые ведутся в контакте с группой GETA (Франция), а также с работами, ведущимися в Малайзии. Как и в Малайзии, предполагается использование математического обеспечения ARIANE, разработанного GETA [65].

Индия. В 1983 начались работы по МП в Индии, в Тамильском университете, имеющие целью создание скромной системы прямого типа, предназначенной для перевода с русского языка на тамильский текстов по астрономии. Система делается на микрокомпьютере [65].

Индонезия. Недавно начались работы по МП в Индонезии. В Agency for the Assessment and Application of Technology разрабатывается полуавтоматическая система для англо-индонезийского перевода. Предполагается участие Индонезии в проекте ODA [65].

Гонконг. Работы по МП начались в Китайском университете Гонконга в 1968 г. Дж. Слокум в [116] называет сделанную там систему CULT (Chinese University Language Translator) наиболее успешной системой машинно-человеческого перевода. С 1975 г. начали использовать для перевода с китайского на английский Acta Mathematica Sinica, а с 1976 г. — Acta Physica Sinica. Неясно, насколько система изменяема для перевода другого типа текстов.

II.5. Проект EUROTRA

Описание работ по МП, ведущихся в Европе, мы начнем с проекта EUROTRA, в котором объединены усилия многих стран. В настоящее время EUROTRA — это крупнейший в мире проект создания переводческих систем, как по числу участвующих стран, так и по числу языков.

В этом проекте сотрудничают свыше 110 специалистов из 16 городов Бельгии, Великобритании, Греции, Дании, Ирландии, Италии, Люксембурга, Нидерландов, Франции, ФРГ. Кроме того, имеется центральная группа и секретариат, находящиеся в Женеве (Швейцария).

EUROTRA является также наиболее амбициозным из всех переводческих проектов по постановке задачи: достижение высококачественного перевода с любого языка на любой в пределах группы языков стран Европейского Экономического Сообщества, т. е. группы, в которую входят английский, голландский, греческий, датский, итальянский, немецкий, французский. Вопрос о присоединении испанского и португальского в 1988 г. еще не был решен окончательно. Предполагается переводить документацию Комиссии ЕЭС, которая по положению Комиссии должна выходить обязательно на языках всех стран сообщества. Это положение делает объем переводческой работы огромным, и до введения средств автоматизации ее выполнение требовало свыше 1800 переводчиков.

В настоящее время Комиссия ЕЭС пользуется системой SYSTRAN, о которой см. в п. II.15.

Работы по проекту EUROTRA начались в 1978 г., предполагалось сначала, что система-прототип будет готова в 1988 г., затем, что в 1990 г., а практически действующая система — в 1993 г. Об истории и общей схеме системы см. [35, 71, 72, 74, 81].

Вся система должна быть типа трансфер. Представление данных на этапе преобразования должно быть единообразным, независимо от входного и выходного языков, так, чтобы через него возможен был переход от любого языка к любому, требующий лишь минимальных структурных преобразований и смены лексики.

Исследовательские группы, работающие в разных странах, должны сделать системы анализа и синтеза для своих языков. Центральная группа должна разработать правила этапа преобразования, с одной стороны, и математическое обеспечение системы, с другой. При этом по замыслу национальные группы не обязаны делать системы анализа или синтеза для разных языков одинаково, за ними должна сохраняться свобода в выборе подходов и решений в пределах некоторой общей схемы.

В общей схеме предполагается, что и в анализе, и в синтезе переход между текстом на некотором языке и представлением уровня трансфер будет многоуровневым. Рассматриваются следующие уровни: 1) уровень текста на естественном языке (быть может стандартизованным), 2) морфологический уровень, 3) уровень составляющих, 4) уровень синтаксических категорий и отношений, 5) уровень семантических категорий и отношений-глубинных падежей (case — frame structure). Этот уровень является уровнем трансфер. Предложенная общая схема обычно упоминается как $\langle C, A \rangle T$ -схема ($\langle C, A \rangle T$ -framework). Здесь A — совокупность атомарных единиц соответствующего уровня, C — грамматика в виде совокупности правил, T — общее преобразование, расчленяемое на T_k — преобразования для перехода от уровня к уровню. Основной формализм преобразований T_k — недетермированное преобразование «дерево — дерево».

Переходы от уровня к уровню должны делаться так, чтобы обеспечивалась стыковка соответствующих переходных представлений как выхода одного уровня и входа другого, независимо от способа их получения с тем, чтобы грамматики и алгоритмы для разных преобразований T_k могли бы разрабатываться независимо.

Стремление обеспечить разным группам специалистов достаточную свободу в выборе стратегии при проведении анализа и синтеза вместе с необходимостью избежать повторов и разнобоя привело создателей EUROTRA к следующему решению. Стратегия анализа или синтеза формулируется в виде специальных правил, управляющих порядком обра-

щения к перерабатываемым и лингвистическим данным, которые формулируются на специальном языке, отличном от того, на каком пишутся правила грамматик. Программное обеспечение системы строится в виде некоторого генератора трансляторов, который на основе спецификаций языков записи грамматических и управляющих правил генерирует транслятор для конкретного варианта анализа или синтеза. С помощью этого транслятора получается внутренняя запись соответствующей грамматики S и программа, реализующая алгоритм преобразований T [71]. Такое решение, по мнению авторов, позволяет избежать необходимости строить универсальный транслятор или, наоборот, вводить жесткие ограничения на допустимые методы. Вместе с тем такое решение безусловно является очень трудновыполнимым.

В 1987 г. был готов малый прототип с ограниченным словарем, который оказался очень медленно работающим: свыше 20 минут на одно предложение. Европейский парламент создал комиссию для рассмотрения состояния дел, которая, указав на много недостатков (в частности, недостаточная разработанность грамматик и словарей для конкретных языков), сочла тем не менее нужным продолжить работы, несмотря на то, что первоначальные сроки создания прототипа и практически действующей системы выдержаны не будут. Комиссия рекомендовала завершить в 1988 г. исследовательскую стадию и начать работу, ориентированную на создание практически действующей системы [65].

II.6. Работы по машинному переводу во Франции

Наиболее активной группой, работающей в области МП во Франции, является несомненно GETA (Groupe d'études pour la traduction automatique) в Гренобльском университете, которая долгое время работала под руководством одного из крупнейших в мире специалистов по МП — профессора Б. Вокуа.

В работах Б. Вокуа не раз формулировались как новые общие подходы и идеи МП, так и новые конкретные решения тех или иных переводческих проблем (см., например, [131—134]).

В 1961—1971 гг. эта группа построила и проверила в обширном эксперименте на текстах свою первую переводческую систему для перевода с русского языка на французский. Она подробно описана в книге Б. Вокуа [131], см. о ней также в [13].

На основе полученного опыта эта группа разработала в 1971—1977 гг. вторую систему, экспериментальная проверка которой началась в 1977 г. Эта вторая система была интерлингвого типа. Экспериментальная проверка этой системы показала, что для получения удовлетворительных результатов при интерлингговом подходе надо уметь сохранять в переходной структуре сведения о поверхностной организации входного текста. В частности, многие погрешности перевода при помощи этой системы были обусловлены следующим обстоятельством. В синтаксическом синтезе выходного (переводящего) текста был установлен жесткий порядок слов, обусловленный только нормами французского языка. Вместе с тем в русском тексте порядок слов, как известно, является гораздо более свободным и, в частности, он является носителем сведений о коммуникативной организации текста (деление на тему и рему или новое и старое и т. п.). Эти сведения при анализе русской фразы не извлекались, а затем утрачивались.

Как уже говорилось выше, этот опыт послужил основанием для широкого распространения систем типа трансфер, где выходная структура более жестко обусловлена входной.

О второй системе русско-французского перевода группы GETA см. [133].

Группа GETA в своих работах всегда уделяла очень много внимания математико-алгоритмическому аспекту переводческих систем.

Для упомянутой выше второй переводческой системы были построены четыре системы: ATEF, SETA, TRANSF, SYGMOR, каждая из которых включает в себя формализм для записи перерабатываемых и лингвистических сведений, алгоритм применения заданной грамматики к перерабатываемым данным и программы, реализующие работу правил, заданных в указанном формате, в соответствии с заданным алгоритмом.

Три из них работали со словарями.

ATEF проводила поиск в словаре русских основ и морфологический анализ, а также строила некоторое фиктивное дерево фразы, в котором узлы всех входных словоформ подчинялись узлу, представлявшему фразу в целом.

TRANSF проводила смену лексики при преобразовании.

SYGMOR реализовывала морфологический синтез словоформ французского языка и строила переводящее предложение.

На долю SETA приходилась основная часть работы, поскольку все этапы процесса перевода, кроме названных, были реализованы как преобразования «дерево — дерево», которое и реализовывалось при помощи SETA. Со временем SETA была заменена на несколько иную систему для преобразования «дерево — дерево» — ROBRA.

Кроме того, группа GETA построила еще диалоговую систему ARIANE, цель которой — облегчить работу лингвиста при создании переводческой системы. ARIANE позволяет в процессе диалога (ответами на ее вопросы) описывать системы признаков, используемых в словарях и правилах, задавать и исправлять грамматики для разных этапов перевода, запускать в работу процесс перевода в целом или отдельные его этапы с помощью каких-то из указанных выше четырех систем, задавать степень детальности отображения хода работы в выдаваемых данных, запоминать те или иные характеристики переводимых текстов или хода работы с ними и многое другое.

Более ранняя версия — ARIANE 78 начала использоваться в Гренобле в 1978 году, а более усовершенствованная — ARIANE 85 появилась в 1985 г. Выше уже упоминалось, что ARIANE используется во многих странах. Описание ARIANE, ATEF, ROBRA, TRANSF и SYGMOR см. в [40, 42], а также в [13].

ARIANE рассчитано на использование лингвистических данных без обращения к базам данных о предметной области и вывода, и на работу в пакетном режиме, т. е. без интерредактирования.

Кроме русско-французского перевода ARIANE использовалось для французско-английского и немецко-французского перевода [133]. Наиболее существенным приложением ARIANE было его использование во Французском национальном проекте Calliore, в рамках которого создавалась система Calliore-Aéro для французско-английского перевода текстов по авиации. В 1986 г. было решено делать также англо-французскую систему для перевода текстов по использованию компьютеров и переработки данных — Calliore-Info. Участие в проекте Calliore стимулировало перенос ARIANE на IBM PC AT, что удобно для разработки систем.

О дальнейшем развитии аппарата преобразования цепочек и деревьев см. [44].

В GETA разработано также программное обеспечение для лексикографических работ ATLAS и для создания словарей VISULEX [36].

Группа GETA всегда была источником интересных новых идей и точек зрения на машинный перевод. Кроме упомянутых выше работ Б. Вокуа, укажем [41, 43, 45].

В последнее время в GETA начались работы по интерактивному переводу. Р. Заяк [140] предложил следующий путь разрешения неоднозначностей при интерактивном переводе с языка пользователя на чужой язык. Установив в некотором месте текста структурную неоднозначность, система задает пользователю вопрос о том, что он хотел сказать, в виде перифраз разных вариантов анализа. Например, встретив сочетание «интересные книги и журналы» система выдаст запрос в виде: «журналы и интересные книги» или «интересные книги и интересные журналы». Для выбора пользователю не нужно ни лингвистических знаний, ни знаний об устройстве системы.

Кроме Гренобля, исследования по МП ведутся также в Париже, Нанси, Пуатье.

Об использовании системы SYSTRAN во Франции см. в п. II.15.

II.7. Работы по машинному переводу в ФРГ

В ФРГ крупнейшая группа, работающая в области МП, находится в Саарбрюкене. В Саарском университете работы по автоматическому синтаксическому анализу начались еще в 1957 г., а работы по переводу — в 60-х годах. Первоначально система SUSY разрабатывалась для русско-немецкого перевода, затем была расширена до многоязычной системы с входными языками: английский, немецкий, русский, французский, эсперанто, — и выходными: английский, немецкий, французский. По типу SUSY — это система типа трансфер.

Группа Саарского университета долгое время работала в обстановке недостаточного финансирования, что позволяло ей сосредоточиться на исследовательском аспекте в отвлечении от проблем реализации. Видимо поэтому система SUSY не получила широкого распространения, хотя и используется для переводов в работах самого университета.

В 1981 г. начались работы над усовершенствованным вариантом SUSY-II.

В 1985 г. в Саарском университете начался проект MARIS (Multilinguale Anwendung von Referenz-Informationssystem) по построению многоязычной информационно-поисковой системы, в рамках которого предусмотрено создание системы частично автоматизированного перевода STS (Saarbrücken Translation Service), которая базируется на исследовательских разработках SUSY.

В Саарском университете ведутся также работы с более далеким прицелом. А. Роткегель работает над системой TEXAN, предназначенной для перевода в паре языков английский — немецкий текстов международных торговых соглашений. При этом учитывается, что такие тексты пишутся с соблюдением определенных ограничений, различающихся в разных языках. Анализ в TEXAN должен давать представление текста в виде ситуаций (в терминах актов и событий с указанием участников, объектов, времени, места и т. д.), а также выявлять коммуникативный статус ситуаций, иллокуции и т. п. [109].

На исследованиях Саарского университета базируется также проект ASCOF (Analyse und Synthese des Französischen mit COMSKEE), где COMSKEE — Computing and String Keeping Language [39].

В ФРГ ведется несколько разработок совместно с Японией.

Группа Саарского университета совместно с группой из университета Киото работает над системой, которая будет переводить заголовки статей с немецкого на японский. Для анализа немецкого используется SUSY, для синтеза японского — TITRAN (см. п. II.2).

Штутгартский университет совместно с японской фирмой Fujitsu разрабатывает систему японо-немецкого перевода, также для перевода

заголовков научных статей. Для анализа используется система ATLAS/II, а для синтеза — система семантического синтеза SEMSYN [108].

В Саарбрюкене, Берлине, Билефельде и Штутгарте имеются группы, участвующие в проекте EUROTRA.

В ФРГ имеется несколько очень крупных терминологических банков, предназначенных для перевода, которые заслуживают упоминания, хотя, как отмечалось выше, при составлении данного обзора не ставилась цель охватить систематически автоматические словари.

Терминологический банк фирмы Сименс носит название TEAM. Фирма Сименс обновляет за пять лет около половины выпускаемой продукции и потому нуждается в постоянном обновлении сопровождающей документации. Объем этой документации составляет до 1 млн. страниц в год, стоимость документации примерно равна стоимости продукции «в металле». Документация переводится на множество языков переводческим бюро, насчитывающим 400 переводчиков. Примерно 20 лет назад, поняв, что следует призвать на помощь ЭВМ, в переводческом бюро начали создавать терминологический банк данных, который сейчас насчитывает около миллиона терминов. Словарная статья термина может содержать до 3000 символов и быть разделена на 99 полей. В них записываются грамматические сведения, сведения об источниках, определения, синонимы, а также способы выражения данного термина на английском, датском, испанском, итальянском, немецком, португальском, русском, французском языках. Имеются также поля для подключения других языков в случае надобности.

Кроме самой фирмы Сименс, с 1981 г. TEAM обслуживает 12 организаций, среди которых и организации вне ФРГ, в том числе ряд издательств. Для издательств готовятся словари по отдельным отраслям, причем от заказа до появления напечатанного словаря проходит не более 6 недель [46].

Предполагается стыковка коммерческой переводческой системы METAL (см. п. II.15) с терминологическим банком TEAM [31].

Другой крупнейший терминологический банк данных принадлежит федеральному переводческому бюро — Bundespachenamт. В 1980 г. этот терминологический банк уже содержал около полутора миллионов входов (887 тыс. англо-немецких, 212 тыс. французско-немецких, 300 тыс. русско-немецких, 19 тыс. — остальные языки). Обслуживающая его система LEXIS позволяет: 1) выдавать словари в печатном виде или на микрофильмы; 2) получать глоссарии по определенной тематике; 3) строить глоссарий к тексту; 4) запоминать переведенные тексты. Кроме того, переводчик может по специальному запросу получать кроме перевода следующие данные: 1) определение смысла термина, 2) примеры употребления термина в контекстах, 3) синонимы и антонимы, 4) отсылки к объемлющим понятиям или, наоборот, к более узким (эти сведения нужны не только переводчикам, но и терминологам для получения семантических полей, словарей понятий и т. д.), 5) отсылки к литературе, где можно найти объяснение, примеры употреблений и т. д. В переводимых текстах 60 % составляют английские, 20 % — французские, 10 % — русские. По тематике — около половины — это инструкции для армии, в основном технические, во второй половине преобладают документы, контракты и т. п. [77].

II.8. Работы по МП в Великобритании

В Великобритании работы по МП начались очень рано. Как известно, сама идея использования ЭВМ для перевода впервые была высказана в переписке английского математика А. Бута с американским лингвистом У. Упвером еще в 1946 г. В период первого подъема интереса к МП в

Англии работали группы в Лондонском университете, в Национальной физической лаборатории в Теддингтоне, а также широко известная работами не только по переводу, но по более широкому кругу проблем вычислительной лингвистики группа Кембриджского университета под руководством М. Мастерман (CLRU — Cambridge Language Research Unit). Несколько позднее начались работы в университете Эдинбурга.

Затем, как и в других странах, последовал спад активности.

В последнее время работы возобновились. Английские специалисты по МП участвуют в проекте EUROTRA (Университет Эссекса, университет Манчестерского Института науки и техники — UMIST — University of Manchester Institute of Science and Technology).

В UMIST ведутся также работы по созданию систем перевода с японского языка и на японский язык для англоязычного пользователя, который не знаком с японским языком.

Одна из них под названием AIDTRANS — человеко-машинная система прямого типа, выдающая пользователю несколько вариантов перевода с японского на английский, с тем, чтобы он мог выбрать или составить правильный, руководствуясь знанием предметной области.

Другая система — NTRAN — это интерактивная трансфер-система для перевода с английского языка на японский. Она допускает на вход только правильные (по ее грамматике) предложения. Если встречается незнакомое слово, то система запрашивает пользователя о его свойствах и формирует в словаре соответствующую словарную статью только без перевода на японский. В японском выходном тексте это слово будет просто переписано, причем пока в латинском алфавите. Со временем разработчики хотят ввести транскрибирование на катагану, считая, что для технических терминов этого часто будет достаточно. Система может также запрашивать пользователя о случаях синтаксической омонимии. Например, она может спросить, какому слову подчинен тот или иной предлог, может просить разобрать конструкцию вида Adj N₁ and N₂, где Adj — прилагательное — может относиться к обоим существительным (N) или только к первому, и т. п. Со временем предполагается добавить выбор японских переводных эквивалентов по английскому пояснению к ним. Система NTRAN реализована на персональных ЭВМ на языке Пролог [84].

Исследовательские работы по созданию формальной модели для перевода ведутся в университете Эдинбурга [37].

В 1984 г. в British Telecom начались работы по переводу устной речи (телефонных разговоров). Ожидается, что система будет готова к середине 90-х годов. Языки: английский, испанский, немецкий, французский, шведский. Работы начаты на корпусе 1000 слов из деловых разговоров. Пока количество нерешенных проблем еще чрезвычайно велико, в том числе таких, как независимость от диктора [65].

II.9. Работы по МП в Нидерландах

В Нидерландах соответствующие работы начались сравнительно поздно.

Группа Института прикладной лингвистики в Утрехте участвует в проекте EUROTRA.

Имеются также два проекта разработки систем типа интерлингва.

В фирме Филипс (Philips Research Laboratories) с 1984 г. ведутся работы над системой Rosetta. Предполагаемые языки — английский, голландский, испанский, первая очередь нацелена на англо-голландский перевод. Система должна быть типа интерлингва. В ней используются грамматики Монтегю, причем разработчики хотят установить соответствие синтаксических правил разных языков через общность смысла и

хотят добиться обратимости правил. Тщательно прорабатываются отдельные проблемы перевода: обработка идиом [112], обработка выражений времени [34], обработка сферы действия отрицания [128] и др.

Другая разработка системы интерлингва началась в 1979 г. в BSO (Buro voor Systeemontwikkeling). Система называется DLT (Distributed Language Translation), она задумана как многоязычная интерактивная система. Ее отличие от всех других переводческих систем — использование Эсперанто в качестве языка-посредника. В конце 1987 г. был продемонстрирован малый прототип с ограниченной грамматикой для англо-французского перевода. Практически действующая система должна быть готова в 1993 г. [136].

II.10. О работах по МП в других Европейских странах

Италия. В пятидесятых годах в Миланском университете активно работала группа под руководством С. Чеккато. В середине шестидесятых годов в информационном центре Евратома в г. Испра (Италия) началось практическое использование системы GAT, сделанной в Джорджтаунском университете (США). Впоследствии система, используемая в Евратоме, несколько расширялась и модифицировалась. Десятилетний опыт работы был обобщен в отчете Б. Достерт [55].

Известны исследования по компьютерной лингвистике, ведущиеся в университете г. Пиза. Институт вычислительной лингвистики университета г. Пиза участвует в проекте EUROTRA.

Швеция. В университете Лунда ведутся работы над экспериментальной многоязычной системой SWETRA [115]. Известны исследования по компьютерной лингвистике группы KVAL, в Стокгольмском университете и университете Линкопинга.

Норвегия. В университете Бергена ведутся работы над системой англо-норвежского перевода ENTRA совместно с фирмой WCC (Weidner), о которой см. ниже п. II.16.

Финляндия. В университете Хельсинки ведутся совместные работы с фирмой IBM в Испании (см. ниже).

Дания, Копенгагенский университет участвует в проекте EUROTRA.

Бельгия. Университеты Льежа и Левена участвуют в проекте EUROTRA [127]. Они же участвуют в работах над системой METAL (см. п. II.15). В университете Брюсселя строится на микрокомпьютере система для перевода руководств по компьютерам с английского на французский [65].

Испания. Работы по МП в Испании начались недавно.

Отделение фирмы Сименс в Барселоне работает над немецко-испанской версией системы METAL [31]. В Барселоне имеется также группа, участвующая в проекте EUROTRA.

Исследовательские центры IBM в Испании и Израиле собираются делать системы для перевода с английского на испанский и иврит для использования в первую очередь в самом IBM. К разработке привлекаются университеты (в частности, университет Хельсинки), предполагается со временем привлечь к работе другие центры IBM в Европе [57, 65].

Швейцария. Институт Семантических и когнитивных исследований в Женеве активно участвует в работах EUROTRA. В Женеве, как указывалось выше, базируется центральная группа EUROTRA и секретариат.

В том же Институте разрабатывается система для перевода в пределах очень ограниченного подязыка: сообщения о требующихся работах. За основу взяты газетные сообщения, которые печатаются на трех языках: итальянском, немецком, французском. Эти сообщения состоят в основном из именных групп, в них мало глаголов, нет придаточных

предложений. Практически нет проблем омографии и полисемии ввиду ограниченности тематики. Система нерасширяема [65].

Греция, Ирландия и Люксембург участвуют в проекте EUROTRA.

Чехословакия. Исследовательская группа Карлова университета в Праге широко известна своими работами в области теоретической и прикладной лингвистики, в первую очередь работами, посвященными коммуникативной организации текстов.

С 60-х годов в Карловом университете ведутся также разработки по машинному переводу, на некоторое время они были свернуты, возобновились в 1976 г. Делается экспериментальная система человеко-машинного перевода с английского на чешский. В 80-х годах началась работа по созданию системы перевода с чешского на русский. Планируется ввод ее в эксплуатацию в 1990 г. [59, 60, 75, 114].

Болгария. В Софии в Институте индустриальной кибернетики и роботизации имеется проект построения системы англо-болгарского перевода [105].

ГДР. В ГДР работы по машинному переводу ведутся давно. Еще в 1963 г. был проведен первый эксперимент по переводу с английского на немецкий, подготовленный группой математической и прикладной лингвистики Германской Академии наук в Берлине [78]. В настоящее время исследовательская группа под руководством Ю. Кунце работает в Центральном лингвистическом институте Академии наук ГДР [79]. В университете Карла Маркса в Лейпциге находится в процессе разработки система «в помощь переводчику» TAS (Translator's Assistant System), снабженная англо-немецким словарем.

Польша. В Политехническом институте г. Жевуша строится система для перевода с русского языка на польский рефератов по химии и биохимии. Планируется ввод в практическую эксплуатацию в 1989 г. [62].

Венгрия. Под руководством Ф. Паппа в Дебреценском университете, а затем в университете Будапешта ведутся обширные работы по построению словарей при помощи ЭВМ [104].

В Академии наук ВНР исследуются лингвистические проблемы, специфические для венгерского языка, решение которых может быть использовано и при построении систем МП [106].

II.11. Работы по МП в СССР

В Советском Союзе работы по машинному переводу начались в 1954 г. в Москве в институте, который ныне — Институт прикладной математики им. М. В. Келдыша АН СССР (ИПМ) и Институте точной механики и вычислительной техники АН СССР (ИТМ и ВТ). Затем довольно скоро возникли работающие коллективы в Ленинграде, Ереване, Тбилиси, Киеве и других городах. Прошли первые конференции по машинному переводу: в 1958 г. в Москве, в 1959 г. — в Ленинграде. В университетах и вузах Москвы, Ленинграда, Киева, Новосибирска возникли отделения, готовившие кадры для работы в области автоматической обработки текстов, их специальность называлась либо математическая, либо прикладная, либо структурная лингвистика. Активное начало работ по МП проходило на фоне широко распространенного увлечения возможностями только что появившейся вычислительной техники и только что вышедшей из опалы кибернетикой.

Решение ALPAC, упомянутое выше, не столь резко сказалось на работах в СССР, как в других странах. Спад интереса к проблеме МП в нашей стране произошел позднее, уже в семидесятых годах. Это связано не только с отголосками ALPAC, но и с рядом других причин. Постепенно выяснилось, насколько трудна проблема, как обманчивы первые успехи, как катастрофически быстро нарастают трудности по

мере продвижения, если ставить цель получения высококачественного перевода. Выявилась явная недостаточность имеющихся лингвистических описаний для решения задач автоматической обработки текстов при помощи ЭВМ. Отрицательно сказались ряд неудачных решений организационного плана. Но, пожалуй, основным сдерживающим фактором была и остается до сих пор недостаточность технической базы, хроническое отставание от передовых стран наших средств вычислительной техники как по количеству, так и по ее возможностям. Это последнее обстоятельство приводило и приводит к тому, что большинство разработок в области МП прошли очень недостаточную проверку в эксперименте, а некоторые вообще остались «на бумаге». Системы, не уступавшие тем, которые делались в других странах, по лингвистическому и математико-алгоритмическому обеспечиванию, но лишенные требуемого программного обеспечения не могли ни превратиться в практически действующие, ни дать достаточный опыт для дальнейшего продвижения в теории.

Это тем более печально, что теоретический уровень пионерских работ по МП в СССР, уровень понимания проблемы, создания методов для ее решения был в большинстве работавших тогда коллективов очень высоким.

Специалистами нашей страны было очень рано понято многое из того, что позднее стало широко распространенным, а иногда даже представляется сейчас в качестве новых идей. Не претендуя на полноту, укажем некоторые примеры, поясняющие сказанное.

Так, среди коллективов нашей страны не было столь повального увлечения грамматиками Хомского, как в других странах. Обманчивое сходство КС-языков с естественными языками не ввело в заблуждение советских специалистов. Достаточно рано было осознано, что эти модели не могут быть адекватными, что в них формализуются не столько существенные аспекты естественного языка как средства коммуникации, сколько средства описания языка. Было понято, что анализ текстов на естественных языках должен делаться не распознавателями, а устойчивыми преобразователями (это подробно разобрано в [13]).

Основная трехфазная схема и многие другие черты наиболее распространенных сейчас систем типа трансфер были очень четко сформулированы в 1966 г. в работе [16]. Этот подход был воплощен в экспериментальных системах англо-русского перевода, сделанной в ЛГУ, и французско-русского перевода (системе ФР-II), сделанной в ИПМ. При этом в системе ФР-II была найдена такая организация алгоритма, которая обеспечивала разумное ограничение переборov и при этом так, что степень сложности переводимого предложения автоматически регулировала число повторений просмотров, благодаря чему простые предложения анализировались быстро [10, 13]. Некоторые другие моменты управления ходом работы также определялись анализируемой фазой, а не тем, что вообще может встретиться в языке (о системе ФР-II см. также ниже).

Достаточно рано были поняты необходимость учета специфики подъязыков, необходимость многоуровневых систем, полезность использования моделей управления предикатов [6], полезность учета специфических особенностей слов в словарях не только в виде наборов пар (признак — значение), но и в виде специальных правил [11, 21].

Уже в 60-х годах были сделаны детальные описания грамматик некоторых языков не в виде правил, предписывающих те или иные акты переработки текстов, а в виде статических условий, которые должны выполняться для правильной синтаксической структуры. На материале русского языка это было сделано Л. Н. Иорданской [7, 8]. На материа-

ле французского языка эти условия были не только сформулированы, но и воплощены в переводческой системе ФР-II [12, 13].

Использование правил предпочтения, которое сейчас получает все более широкое распространение, было предложено применительно к синтезу текстов еще в 1965 г. И. А. Мельчуком [22], а в применении к анализу разрабатывалось с разными подходами Г. С. Цейтиным [28] и автором [15].

К сожалению, этот богатый идейный задел не получил должного применения, когда наметился переход к практической реализации перевода.

В подавляющем большинстве системы МП в СССР строились для перевода в паре языков в направлении «с чужого на свой». Исключением среди ранних работ являются разработки, которые велись в Ереване, где делался перевод как с армянского на русский, так и с русского на армянский. Чаще всего разрабатывались системы для перевода с английского на русский: экспериментальные системы для этой пары языков были сделаны в Москве (ИТМ и ВТ, ИПМ, ЦНИИПИ, КБПА, позднее Информэлектро, ВЦП), в Ленинграде (ЛГУ) и в других местах*). Системы французско-русского перевода были сделаны в ИПМ, позднее — в Информэлектро, ВЦП. Система немецко-русского перевода — в ВЦП. В Тбилиси велись работы по русско-грузинскому переводу, в Киеве — по русско-украинскому. Были разработки небольшого объема и для других пар языков.

Из числа экспериментальных советских систем наиболее полную проверку в экспериментах по переводу прошла система ФР-II (ИПМ). Работы по машинному переводу с французского языка на русский велись в ИПМ в течении ряда лет. Были построены две системы: система ФР-I прямого типа (1954—1961 гг.) и затем система ФР-II типа трансфер. Система ФР-II была системой с множественным синтаксическим анализом фильтрового типа, с преобразованием от входного языка к выходному на уровне синтаксических структур, имеющих вид деревьев зависимостей. Для нее был подробно разработан аппарат преобразований «дерево — дерево» [9]. В ней впервые была построена подробная французская грамматика фильтрового типа для переводческой системы. Фильтры — это ограничения на сочетаемость синтаксических связей в правильно построенной синтаксической структуре, о которых упоминалось выше. Алгоритм анализа в системе ФР-II существенно отличался не только от широко распространенных тогда алгоритмов анализа по составляющим снизу вверх, ориентированных на грамматики Хомского, но и от предложенного для фильтрового подхода И. Лесерфом [83] и использованного в ЛГУ [5]. Кроме тех особенностей, которые уже упоминались выше (ограничение переборков, самонастройка), специфика алгоритма анализа ФР-II состояла в применении фильтров к набору гипотетических синтаксических связей до его расщепления на отдельные деревья с целью сокращения этого набора. Сокращение шло циклически и существенно опиралось на принцип предпочтения уникальных связей (см. [13]), а число циклов, как было сказано, регулировалось сложностью анализируемого предложения. Эти особенности сделали фильтровый метод компьютерно реализуемым.

Позднее сходный алгоритмический подход был применен в системах Информэлектро: ЭТАП-1 для французско-русского перевода и ЭТАП-2 для англо-русского перевода.

*) Из не расшифрованных ранее здесь употреблены сокращения: ЦНИИПИ — Центральный научно-исследовательский патентный институт, КБПА — конструкторское бюро промышленной автоматики, ЛГУ — Ленинградский государственный университет, о ВЦП см. ниже.

Подробное описание системы ФР-II и обзор систем, сделанных в СССР в 50-х — 70-х годах, можно найти в [13]. Описание систем ЭТАП содержится в [1, 2, 29].

Переход к практической реализации машинного перевода в СССР складывался следующим образом. В 1973 г. во Всесоюзном центре перевода научно-технической литературы и документации ГКНТ и АН СССР (ВЦП) был создан отдел машинного перевода, которому была поручена практическая реализация МП. К сожалению, в ВЦП не был осознан и применен тот существенный теоретический задел, который уже существовал к этому времени.

Для практического использования была взята система прямого типа для англо-русского перевода — система АМПАР [19]. Того же типа была и сделанная в ВЦП система немецко-русского перевода НЕРПА [20]. Впоследствии эти две системы были объединены в единую систему, названную АНРАП (англо/немецко-русский автоматический перевод), с более усовершенствованным математико-программным обеспечением [26]. Система АНРАП работает на ЭВМ ЕС 1035, скорость — до 4 печатных листов в час. Словари для системы АНРАП содержат 40 тыс. английских входов, 30 тыс. немецких, 50 тыс. русских. В [103] сказано, что с ее помощью в 1986—87 гг. были переведены тексты общим объемом 800 млн. знаков, что составляет около 10 тыс. печатных листов в год. Сопоставляя эти данные с объемом переводов в ВЦП (например, в [27] сказано, что ВЦП перевело в 1984 г. 75 тыс. а.л., а с тех пор объемы переводов возросли), можно установить, что с помощью АНРАП переводится 12—15 % переводимых ВЦП текстов. Данных о качестве продукта АНРАП и о том, насколько ее использование ускоряет работы переводчиков, не приводилось. Результаты работы АНРАП редактируются переводчиками ВЦП. В последние годы под руководством Б. В. Тихомирова разработаны средства редактирования, которые в 1988 г. переносились на персональные ЭВМ, без чего трудно представить себе эффективно действующую переводческую систему [26, 103].

Еще в семидесятых годах в ВЦП велась разработка более современной системы англо-русского перевода [17], однако пока о переходе ВЦП к системам типа трансфер не сообщалось. Не вышла пока на практическое использование и система французско-русского перевода ФРАП, которая делалась в ВЦП [103]; описание ФРАП дано в [24].

Несколько организаций сообщали о вводе в практическую эксплуатацию систем, дающих подстрочник или близкий к нему пословно-пооборотный перевод. В Информэлектро работает пословный перевод заголовков статей по электротехнике в библиографических базах данных с английского и французского языков на русский. В Чимкентском педагогическом институте сделан практический пословно-пооборотный перевод текстов по вычислительным машинам с английского на русский. В Ленинградском государственном педагогическом институте им. Герцена сделаны для практического использования пословно-пооборотного перевода версии для перевода с английского и французского языков на русский текстов политического характера. Там же находятся в стадии экспериментальной проверки испано-русская и русско-английская версии [103].

Известны еще следующие экспериментальные разработки.

В Киеве в ВНИИ нефтехим с 1981 г. находится в стадии экспериментальной эксплуатации интерактивная система перевода с английского на русский СИМПАР для текстов по нефтехимии и термодинамики, реализованная на ЕС 1040 на языках ПЛ/1 и Ассемблер [4].

В Харькове в ВНИИТэлектромаш — экспериментальная система перевода с английского на русский для текстов по порошковой металлургии. Словарь 7 тысяч единиц. Реализована на ЕС на ПЛ/1. Дает

подстрочник с устранением омографии по линейному контексту, перестановку рядом стоящих существительных и делает морфологический синтез [3].

Возвращаясь к работам ВЦП, отметим еще следующие разработки. Построен многоязычный словарь автоматический словарь для языков английского, немецкого, французского и русского, который содержит слова общей лексики, а также термины по вычислительным машинам программированию и радиозлектронике. Общий объем — 100 тысяч входов. Ведутся работы по созданию так называемого автоматизированного рабочего места переводчика (АРМ) на базе ЕС 1841, предназначенного для использования как переводчиками, так и редакторами, терминологами, лексикографами и администраторами. Основу разрабатываемого АРМ составляют три подсистемы: 1) подсистема ввода текста, поиска в словарях, редактирования перевода; 2) подсистема поддержки словарей; 3) подсистема создания персональных глоссариев новых терминов [103]. (О других подходах к построению автоматизированных рабочих мест переводчиков см. п. II.17.)

В ВЦП известны также две разработки более перспективного характера. Совместно с Институтом востоковедения АН СССР ведутся работы по автоматизации японо-русского перевода. В настоящее время находится в стадии экспериментального опробования упрощенный японский анализ, который содержит правила расчленения текста на слова (японский текст пишется без пробелов) и некоторые правила анализа линейного контекста, с помощью которых можно иногда установить сведения о форме русского слова (часто неоднозначно: например, падеж дательный или предложный) и об их расстановке. Общая схема работы и правила для анализа японского делаются Э. М. Шаляпиной [30], словарь, русский морфологический синтез и программное обеспечение — ВЦП [103].

Другая разработка исследовательского плана — создание мета-системы для МП ФЛОРЕАТ, которая делается под руководством Ю. С. Мартеньянова на базе разработанной им ранее системы трансформационного синтеза ФЛОС [18].

Что же касается исследовательских советских коллективов, которые начинали работы по МП и успешно развивали их в течение ряда лет, то они постепенно переключились на другую тематику. Например, группа ЛГУ перешла на построение систем понимания текстов в русле работ по искусственному интеллекту и информационно-поисковых систем. Группа ИМП — переключилась на разработку методов алгоритмизации анализа текстов на основе широкого использования не только фильтров, но и предпочтений (на материале русского языка) [14, 15]. Группа, работавшая в Информэлектро, перешла в Институт проблем передачи информации АН СССР и разрабатывает лингвистическое обеспечение для экспертных систем. Отошли от переводческой тематики коллективы, работавшие в Ереване и Тбилиси.

II.12. Работы по МП в Канаде

В Канаде отношение к МП во многом определяется тем, что это страна с двумя государственными языками: английским и французским. Естественно, что большинство разработок ориентировано на эту пару языков. Исследования по МП и разработка систем ведутся в университетах Монреалья, Оттавы, Торонто и в фирмах.

Начиная с 1977 г. в Канаде регулярно эксплуатируется переводческая система МЕТЕО, разработанная в Монреальском университете группой ТАУМ (Traduction Automatique) в 1975—76 гг. Она используется в Бюро переводов Министерства окружающей среды для перевода

текстов метеосводок, поступающих из разных пунктов. Ежедневно переводится до 5 тыс. сообщений. Перевод ведется с английского на французский [119].

МЕТЕО — это система типа трансфер с небольшим словарем (около 1500 слов) и довольно жесткой грамматикой. Система переводит только те входные предложения, которые признаются правильными ее грамматикой, а те, в которых есть какая-либо неправильность, в частности, из-за ошибки ввода, передаются для перевода человеку. Первоначально процент таких отдаваемых человеку предложений был очень высок — порядка половины от общего числа, постепенно он был сведен к одной десятой. Принятые системой предложения переводятся так, что уже не требуют редактирования (редкая ситуация в МП).

Во второй половине 80-х годов канадские исследователи начали работать над системой, которая должна сменить МЕТЕО, и которая будет уже не переводить, а синтезировать английские и французские тексты на основе некоторого формализованного описания метеобстановки (температуры, влажности, скорости и направления ветра и др.). Синтез состоит из двух этапов. Первый не зависит от языка и состоит в получении по заданным сведениям некоторого логического представления ситуации. Второй этап — это синтез текста по такому описанию. Он уже ведется с привлечением словарных и грамматических данных о соответствующем естественном языке [76].

После МЕТЕО тот же коллектив был занят некоторое время разработкой системы TAUM — AVIATION, которая должна была вести англо-французский перевод текстов технологических инструкций к самолетам (1977—1981 гг.). Однако выяснилось, что разработка системы обходится дорого, степень сложности текстов несравненно выше, чем в МЕТЕО, в то время как даже франкоговорящие специалисты по авиации достаточно хорошо понимают английский текст. Поэтому продолжение этих работ было признано нецелесообразным [67].

Затем группа под руководством П. Изабель занималась разработкой этапа трансфер [68]. В настоящее время, уже в Канадском исследовательском центре по информатизации работы (Centre Canadien de Recherches sur l'Informatisation de Travail) она разрабатывает еще одну узко специальную систему: CRITTER. Эта система предназначена для перевода в паре языков английский — французский в обе стороны еженедельных отчетов Министерства сельского хозяйства о состоянии рынка сельскохозяйственной продукции.

CRITTER относится к типу трансфер. В ней имеется два уровня представлений для переводимого текста: поверхностно-синтаксическая структура в виде размеченного дерева и семантическое представление в виде аргументно-предикатной структуры (направленный ациклический граф). Синтаксические правила состоят из декларативной части, единой для анализа и синтеза, и указаний, управляющих ходом процесса. К середине 1988 г. был готов экспериментальный словарь с 500 единиц, для английского и французского была описана морфология, синтаксическая часть разрабатывалась, в конце года ожидался переход к стадии получения первых экспериментальных переводов [69].

Компания SOCATRA (Société Canadienne de Traduction) в Монреале сообщила о завершении в 1988 г. системы англо-французского человеко-машинного перевода — XTL. С помощью этой системы SOCATRA ведет переводы по заказу, причем заказчик может получать как неотредактированный продукт, так и перевод, окончательно отредактированный переводчиком компании [65].

Из интересных разработок исследовательского плана следует указать систему GOSSIP (Generation of Operating System Summaries in Prolog), которая делается в Odyssee Recherches Appliquées Inc.

(Montréal). Она предназначена для синтеза английских текстов и опирается на известную модель «смысл — текст» И. А. Мельчука [23]. Синтез текста в ней многоуровневый: по исходной информации строится концептуально-коммуникативное представление, оно преобразуется в семантическое, то — в синтаксическое, которое в свою очередь преобразуется в морфологическое, а последнее — в текст. В системе GOSSIP используются правила перифразирования и аппарат лексических функций. Система реализована на Прологе и в середине 1988 г. находилась в стадии первых экспериментов [66].

II.13. Работы по МП в США

Выше уже говорилось о том, что первоначально МП развивался особенно активно в США. Вскоре после Джорджтаунского эксперимента 1954 года в исследования по МП включились 18 университетов США, ряд фирм. После ALPAC положение резко изменилось, а при втором подъеме работы по МП в США не приобрели того размаха, который они имеют в Японии.

Первый проект по МП возник в 1952 г. в Джорджтаунском университете. Он-то и дал Джорджтаунский эксперимент, а затем после десяти лет работы в 1964 г. вышла на практическое использование система GAT (Georgetown Automatic Translation) для перевода с русского языка на английский. Она использовалась в Комиссии по атомной энергии (США) и в информационном центре Евромата в г. Испра (Италия). В 1973 г. Б. Достерт выпустила подробный отчет об использовании GAT и об оценке ее работы пользователями. Система GAT относилась к числу прямых, перевод был близок к пословному. Однако, несмотря на низкое качество результатов, она использовалась в Евратоме до 1976 г., а в Комиссии по атомной энергии — до 1979 г. Как указывает Б. Достерт, подавляющее число пользователей совершенно не знало русского языка и не могло быстро получить переводы другим способом, тогда как результаты GAT позволяли получить общее представление о содержании переводимых научных статей [55].

Впоследствии систему GAT сменила система SYSTRAN (см. п. II.15).

В шестидесятые годы в Бюро иностранной технологии ВВС США работала система русско-английского перевода MARK II, которую в 1970 г. сменила та же система SYSTRAN.

О SYSTRAN и других коммерческих системах, сделанных в США, но получивших распространение и в других странах, будет сказано в пп. II.15, 16. Здесь мы коснемся исследовательских и экспериментальных работ. Дж. Слокум (Microelectronics and Computer Technology Corporation, Austin) в своем обзоре МП 1985 г. [116] пишет: «Основная проблема, связанная с исследовательскими МП системами — это их отсутствие (nonexistence) в США» (с. 15). Однако У. Хатчинс в неоднократно упоминавшемся обзоре [65] отмечает: «В последние годы возрастало число экспериментальных проектов по МП в США» (с. 23). У. Хатчинс называет следующие.

Кольгейтский университет — интерлингвовая система TRANSLATOR с использованием экстралингвистических знаний и знаний о дискурсе.

В университете Карнеги-Меллон (CMU) есть два проекта. Один из них делается под руководством Дж. Карбоннеля и М. Томита. В нем предполагается использовать унификационные грамматики М. Кея для представления знаний о языках (сначала — английском и японском, позднее испанском, итальянском, немецком, французском), а также концептуальные данные об определенной предметной области (общение доктора с пациентом) [120]. В другом проекте CMU цель работы — интерактивный диалог.

Кроме того, Центр МП этого университета создает объединенный международный проект, к которому привлекаются IBM, Хьюлетт — Паккард и ряд японских компаний по вычислительной технике.

В университете Нью-Мексико имеется экспериментальная система англо-китайского перевода XTRA (руководитель — X. Huang).

В университете Калифорнии ведутся эксперименты с японо-английской системой JETR (руководитель — R. Yoshii).

В университете Британской Колумбии ведутся эксперименты по синтаксическому анализу (руководитель — R. Sharp).

В университете Техаса, где в свое время была создана система METAL (см. п. II.15), ведутся работы на материале английского и китайского языков по созданию правил, которые могут быть использованы как для анализа, так и для синтеза [70].

В Джорджтаунском университете возобновляются работы по МП.

В университете Джонса Хопкинса делается система прямого перевода TRANSOFT для перевода с немецкого на английский текстов по медицине. Работа ведется совместно с Массачузетской клинической больницей и Фрайбургским университетом.

Исследовательские работы ведутся также в Йельском университете, Технологическом институте Джорджии, в некоторых фирмах.

В нескольких компаниях, в том числе Linguistic Products, ТП (Telecommunications Industries Incorporation) разрабатываются средства помощи переводчику. См. также п. II.16 и II.17.

II.14. Многоязычные проекты в Южной Америке

У. Хатчинс [65] сообщает о двух крупных многоязычных проектах, имеющих в странах Южной Америки.

Работы в рамках CADA (Computer Assisted Dialect Adaptation) ведутся под руководством Д. Вебера с 1979 г. Этот проект существенно опирается на близость учитываемых языков. Это скорее помощь человеку в переводе, чем собственно МП. В нем имеются следующие этапы: морфологический анализ (выделение основ, установление грамматических характеристик суффиксов), лексический трансфер, использующий словарь билингвых соответствий основ, морфологический синтез. Работы по подготовке лингвистического обеспечения охватывают языки разных стран: Перу (кечуа, кампа), Колумбии (туканоан), Гватемалы (какчикель), Филиппин, Эквадора (кичуа), Бразилии (тупи). Программное обеспечение делается на языке Си для микрокомпьютеров DEC.

Другой проект выдвинут в Боливии и носит название ATAMIRI, его руководителем является I. Guzman de Rojas. Строится многоязычная интерактивная интерлингвая система, в которой в качестве языка-посредника выбран язык аимара (Aymara) — один из индейских языков Южной Америки. ATAMIRI — это, с одной стороны, «переводчик» на языке аимара, а, с другой стороны, сокращения для Automata Traductor Algoritmico Multilingue Interactivo Recursivo Inteligente. Гузман де Ройяс утверждает, что благодаря регулярности морфологии и синтаксиса язык аимара может быть идеальным языком-посредником для многоязычной системы. Сейчас пока система чисто синтаксическая, причем для представления перерабатываемых сведений используются матричные, а не древесные структуры. Система рассчитана на интерактивное участие человека в процессе перевода. Имеется сообщение, что ATAMIRI используется в переводческом центре в Панаме для перевода с английского на испанский. Она опробовалась также для перевода с английского на голландский, немецкий и французский, а в 1989 г. планируется добавить к выходным языкам итальянский и шведский [58].

II.15. Коммерческие МП-системы

В этом пункте речь идет о коммерческих системах с пред- и постредактированием. Они делались в США, используются в разных странах.

SYSTRAN. Система SYSTRAN является сейчас наиболее широко распространенной в мире. Она появилась на базе системы GAT (см. выше) в результате разработки, осуществленной под руководством П. Тома. По сравнению с первоначальным вариантом усовершенствована в смысле модульности и использования глоссариев для узких предметных областей. Как и GAT, это система прямого типа.

Первой парой языков в SYSTRAN была пара: русский — английский. Их словари содержат свыше 200 тысяч слов общей лексики и свыше 200 тысяч терминов по различным отраслям техники.

SYSTRAN для русско-английского перевода сменила в 1970 г. систему MARK II в BBC США, а в 1976 г. — систему GAT в Евратоме. В BBC США в конце 80-х годов с помощью SYSTRAN переводилось до 100 тысяч страниц в год: советские патенты и технические журналы, в первую очередь, относящиеся к авиации и космосу. Со временем в BBC США стали использовать также немецко-английскую и француско-английскую версии, а также разрабатываются грамматики для итальянского, испанского, португальского и японского в качестве входных языков.

В 1976 г. Комиссия ЕЭС приобрела англо-французскую версию SYSTRAN и хотя словарь оказался неудовлетворительным, но потребность в переводе столь велика, что Комиссия пошла на работы по усовершенствованию словарей и в 1978 г. приобрела еще французско-английскую версию, а в 1979 г. — англо-итальянскую. Словари содержат по 100—200 тысяч единиц. Со временем подключились версии для перевода с английского на голландский, испанский, немецкий, португальский, с французского на голландский и немецкий. Перевод на английский и французский получается лучше, чем на немецкий и голландский из-за специфики порядка слов. В целом результаты пока оставляют желать много лучшего, и в переводах Комиссии переводы, сделанные с помощью SYSTRAN, составляют 2 %.

Фирма Хегох (США) использует SYSTRAN для перевода технических руководств, которые пишутся на упрощенном английском языке. До 50 тысяч страниц в год переводится с английского на испанский, итальянский и французский. Благодаря ограничениям на строение входных текстов фирма получила (по ее сообщениям) за счет ввода SYSTRAN пятикратное увеличение производительности труда переводчиков. В других переводческих бюро, где нет ограничений на вид входа обычно отмечалось ускорение в 2—3 раза [110, 113].

В Канаде фирма General Motors использует англо-французскую и англо-немецкую версии SYSTRAN.

Во Франции Aérospatiale переводит с помощью SYSTRAN авиационные руководства в паре языков английский-французский в обе стороны. Через сеть Minitel во Франции к SYSTRAN подключено 4,5 миллиона пользователей, которые применяют англо-французскую, англо-испанскую, англо-португальскую и испано-английскую версии SYSTRAN в основном для переписки в сфере международного бизнеса.

Другие пользователи разных версий SYSTRAN: штаб НАТО в Брюсселе, Немецкое национальное управление железными дорогами (Bundesbahn) и Исследовательский атомный центр в Карлсруе (Kernforschungszentrum) в ФРГ и другие.

В настоящее время SYSTRAN является почти целиком собственностью компании Гашо (GACHOT S. A. Compaу), которая объединила

интересы всех компаний США и Европы, ранее владевших SYSTRAN. За пределами компании Гашо осталась лишь IONA Company (Япония), которая владеет SYSTRAN Corporation of Japan. Гашо стремится обеспечить единообразие разных версий SYSTRAN, которые разрабатывались постепенно в разных местах. В настоящее время она может предложить 15 версий для разных пар языков: с английского на голландский, испанский, итальянский, немецкий, португальский, русский, французский; японский; с французского на английский, голландский, немецкий; с испанского, немецкого, русского и японского на английский. Разрабатываются также версии для перевода с английского на арабский, с немецкого на испанский, португальский, французский; с китайского, итальянского и португальского на английский язык.

Во второй половине восьмидесятых годов SYSTRAN стали использовать в ряде переводческих бюро: Гашо в Париже, European Center for Automatic Translation в Люксембурге, Бюро Мендес в Брюсселе, CSATA в Италии.

LOGOS. Работа над системой LOGOS началась в 1964 г., в 1971—73 гг. ВВС США применяли ее для перевода военных инструкций с английского на вьетнамский.

В 1978 г. фирма Сименс начала финансировать создание на основе разработок LOGOS немецко-английской системы, которая была сделана через три года, но фирмой Сименс не использовалась. LOGOS заключила соглашение с фирмой Wang, которая перевела систему на свои микрокомпьютеры (первоначально она была реализована на машине IBM). С 1984 г. на рынке еще англо-немецкая версия. Обе эти системы — «гибридные»: с выделенным анализом и объединенным преобразованием и синтезом. Основной двуязычный словарь содержит 100 тысяч входов, он может пополняться словарем пользователя. Покупатели — в основном фирмы ФРГ (Nixdorf, Hewlett — Packard и др.). Есть сообщения о работе над другими версиями.

SPANAM и ENGSPAN — это две системы, сделанные в РАНО (Pan American Health Organisation) в Вашингтоне.

РАНО начала работы по МП в 1975 г., в основном опираясь на подходы GAT. Официальные языки этой организации: английский, испанский, португальский, французский. Для начала была взята пара испанский-английский.

SPANAM — система прямого типа для испано-английского перевода, она начала эксплуатироваться в 1980 г. В 1981 г. началась разработка системы англо-испанского перевода ENGSPAN, которая вступила в строй в 1985 г. Эта система уже типа трансфер. Обе системы предназначены для перевода определенного типа текстов: медицинской документации. Словари (около 50 тыс. единиц) содержат медицинские термины, среди которых допускаются очень длинные: до 25 слов [130].

Эти системы в строгом смысле слова не являются коммерческими, поскольку они используются только самой РАНО. Однако как практически действующие системы, обслуживающие интернациональную организацию, они отнесены к этому разделу.

METAL. Система METAL разрабатывалась в Техасском университете для фирмы Сименс, первоначально для пары языков английский — немецкий (в обе стороны). В 1982—84 гг. она проходила экспериментальную проверку, а в 1985 г. вышла на рынок под названием LITRAS [38, 107, 117].

METAL — наиболее сложная система из числа коммерческих, относится к типу трансфер. Немецкий анализ в ней делается на основе КС-грамматики, английский — на основе расширенной НС-грамматики (GPSG). Анализ последовательно отделен от преобразования. Преобразование включает правила синтаксической перестройки и смены лексики.

Работа тех и других совершается за один просмотр дерева, полученного при анализе, сверху вниз, причем порядок применения правил может регулироваться специальными указаниями. Если дерево анализа не было построено, то перевод идет по составляющим, в сумме покрывающим все предложения.

Система METAL по ориентации многоязычная, организованная так, чтобы языки могли сочетаться друг с другом произвольно. Сейчас разрабатываются грамматики для голландского, испанского и французского языков. В этих работах, кроме Техасского университета, участвуют университеты Левена (Бельгия) и Барселоны (Испания).

METAL реализована на Symbolics Lisp машине, рассчитана на работу в пакетном режиме. Редактирование делается на РС, причем допускается и пред- и постредактирование.

Предполагается подключение METAL к терминологическому банку TEAM фирмы Сименс.

II.16. Коммерческие МЧП-системы

В п. I.1 мы назвали МЧП-системами (т. е. системами машинно-человеческого перевода) системы с интерредактированием, т. е. системы, рассчитанные на участие человека на промежуточных стадиях процесса перевода.

Следует отметить, что в литературе имеется некоторый разнобой. Перевод с интерредактированием часто называется Machine-aided/assisted translation (MAT), хотя иногда его же называют Human-aided/assisted machine translation (НАМТ), оставляя название Machine-aided/assisted human translation (МАНТ) за системами, где переводит человек, использующий автоматический словарь. Представляется, что различие НАМТ (машинно-человеческий перевод) и МАНТ (человеко-машинный перевод) лучше отражает специфику организации работы в каждом случае.

Интерредактирование возможно на всех стадиях работы переводческой системы: при анализе (главным образом для снятия омонимии слов и конструкций), при преобразовании (особенно для выбора переводных эквивалентов), при синтезе (для получения гибкого результата, отражающего особенности коммуникативной организации оригинала). Надо отметить, что пока интерредактирование чаще включается на стадиях анализа и преобразования, чем при синтезе.

В Университете Брайяма Янга г. Прово (США) в течение долгого времени (с 1970 по 1979 гг.) велись работы под руководством А. Мелби по созданию интерактивной переводческой системы. В ней предполагалось иметь один входной язык — английский и много выходных. Сначала считалось, что это позволит получать хорошее внутреннее представление анализируемого предложения, на основе которого можно будет затем синтезировать переводы на разные языки уже без вмешательства человека [86, 87]. Эти надежды не полностью оправдались, так как оказалось, что вмешательство нужно и на других стадиях; кроме того работа интерредактора требовала очень высокой квалификации: детального понимания устройства системы и хода ее работы; система в целом работала медленно. В результате данная система не перешла в ранг практически используемых, но на ее основе возникли две дочерние системы в фирмах, куда перешли разработчики из университета Брайяма Янга.

Одна из этих фирм — Weidner Communication Corporation — теперь лидер среди производителей интерактивных систем на микрокомпьютерах, ныне она называется World Communications Center — WCC.

В Weidner перешли работать в основном математики и программисты.

сты университета Брайяма Янга. С 1980 г. фирма продает переводческие системы. По данным 1980 г. [129] Weidner поставляла за 130 тысяч фунтов стерлингов следующий комплект: 1) микрокомпьютер Digital; 2) 4 терминала; 3) математическое обеспечение для переводческой системы и системы редактирования (text — processing); 4) словарь с 10 тыс. словарных статей. Лингвистическое обеспечение для дополнительной пары языков делалось за 75 тыс.

С 1988 г. фирма WCC принадлежит японскому бюро переводов Bravis.

Интерактивные системы WCC существуют в двух версиях: MacroCAT — система на DEC Vax II и MicroCAT — система на IBM PC/XT, AT.

MicroCAT-система существует в следующих вариантах: перевод с английского на испанский, итальянский, немецкий, португальский, французский и японский; с испанского, французского и японского на английский. Ведутся работы с арабским, китайским, корейским и норвежским языками.

Первые варианты системы были прямого типа, сейчас начинают привлекаться более сложные лингвистические модели (например, LFG — lexical functional grammar).

В Японии MicroCAT продается под названием Micro Pack причем наиболее активно раскупаются версии для японо-английского и англо-японского перевода. К 1985 г. они были проданы в 1900 организаций, а все остальные — в 100 организаций; к 1988 г. было продано уже 3000 комплектов для перевода в паре японский — английский [65].

Переводческая система WCC демонстрировалась в 1986 г. в Москве на выставке Япония 86.

Другая часть группы, работавшей ранее в университете Брайяма Янга (в основном лингвисты) перешла в фирму ALPS.

Для IBM PC/XT ALPS продает три пакета программ.

Transactive — интерактивный перевод с английского на испанский, итальянский, немецкий, французский, с французского на английский.

Autoterm — поиск терминов в языках: английском, голландском, датском, испанском, итальянском, норвежском, португальском, французском.

Translation support system — многоязычный процессор для лексикографических работ, подсчета частот слов, составления конкордансов и т. п.

Разрабатывается ABC-word, который должен давать доступ к словарю Webster, тезаурусам и др.

В последние годы пакеты ALPS приобрели многие переводческие бюро [65].

II.17. Автоматизированное рабочее место переводчика

Переход к широкому практическому использованию переводческих систем заставил по-новому взглянуть на взаимоотношения человека и машины в процессе перевода: осмыслить их не только с точки зрения эффективности использования системы, но и с учетом удобства работы переводчика. Это связано с тем, что объем необходимых исправлений в результатах, выдаваемых машиной оказался большим. Встал вопрос о том, как организовать дело так, чтобы у человека не возникало ощущения, что он только обслуживает систему, вынужден приспособливаться к ней, исправляя ошибки, иногда такие, каких не сделал бы и школьник.

А. Мелби [88, 82] предложил в качестве решения организовать автоматизированное рабочее место переводчика — АРМ (иначе — рабочую станцию — workstation for translator) с тремя разными уровнями учас-

тия машины. АРМ базируется на персональной ЭВМ, которая может быть подключена к сети ЭВМ.

Первый уровень — простейший: автоматизируется обращение к словарям и внесение редакторской правки. Этот уровень не требует, чтобы весь текст был в памяти машины.

При втором уровне весь текст должен быть на машинно-читаемых носителях, а система должна в дополнение к возможностям первого уровня уметь производить определенные операции над текстом в целом, например, собрать все случаи появления в тексте некоторого термина; на этом уровне подключаются программы морфологии и др.

Третий уровень включает переводческую систему, которая может предложить перевод предложения. При этом желательно, чтобы переводческая система обладала встроенными критериями оценки качества своей работы так, чтобы переводчик мог по желанию поручить ей выдавать только те результаты, которые оценены не ниже указанного им порога.

Самое основное состоит в обеспечении возможности простого и легкого перехода от уровня к уровню по желанию переводчика так, чтобы он мог свободно выбирать способ работы и полностью контролировать ситуацию. Такая организация позволяет не требовать, чтобы переводческая система переводила все предложения текста, с ее помощью будут переводиться только те предложения, с которыми она может успешно справиться. Остальное будет переводить человек, переключаясь на более низкие уровни.

Указанные выше пакеты программ ALPS близки к тому, что предложено А. Мелби.

II.18. Комбинированные системы

В некоторых случаях системы перевода объединяются с другими системами обработки текстов. Мы уже указывали случаи использования перевода в информационно-поисковых системах (в основном переводов заголовков статей).

Другой случай объединения представляет фирма Smart Communication, которая имеет два основных продукта: Smart Expert Editor и Smart Translator. С помощью первого составляются технические отчеты (предметная область — химия). Поскольку они получаются на ограниченном языке с известной грамматикой, они легко поддаются переводу. Имеются версии для перевода с английского на испанский, итальянский, португальский, французский и с этих языков — на английский. Системы Smart работают в 30 разных компаниях, крупнейший пользователь — Канадское министерство по трудоустройству и иммиграции [65].

Заключение

Мы осветили, вынужденно кратко, разные типы переводческих систем, от очень сложных — с элементами понимания, т. е. извлечения компонент смысла, до самых простых, автоматизирующих только простейшие операции в переводческом процессе. Возможно, что какие-то переводческие системы остались за рамками данного обзора из-за отсутствия сведений о них. Но уже из приведенного перечня стран, где идут работы по МП, организаций и разработанных систем видно, что работы по автоматизации перевода имеют очень большой размах и продолжают развиваться. Сейчас уже не приходится опасаться нового ALPAC. Неудача и даже провал отдельного проекта уже не останавливают развитие области в целом.

Естественно ожидать, что дальнейшее развитие в исследовательской части будет идти как «вглубь», так и «вширь». «Вглубь» — к представлениям все больше отражающим смысловое, не зависящее от языка, со-

держание переводимого сообщения. «Вширь» — раздвигая рамки отдельного предложения, переходя к абзацу, параграфу, главе, тексту.

В плане практической реализации можно ожидать особенно активного развития интерактивных систем, наряду с повсеместным переходом на трансфер-системы для систем, работающих в пакетном режиме.

Спорить на тему «могут или не могут переводить машины» теперь уже бессмысленно. Ясно, что машины пока не заменили человека-переводчика, как мечталось в начале работ по МП. Но так же ясно, что переводческие системы стали надежными помощниками человека в переводческой работе. При этом одни системы лучше справляются с одной частью переводческих проблем, другие — с другой. Важно уметь для каждой системы четко и честно оценить, что именно она может, в чем ее сильные и слабые стороны. Это нужно как для того, чтобы не вводить в заблуждение возможных пользователей, так и для того, чтобы установить для системы тот способ сочетания ее работы с работой человека, при котором она будет оказывать человеку максимальную помощь.

Количество нерешенных проблем еще очень и очень велико. Хотя теперь МП уже имеет богатую сферу приложений, он не свелся к прикладным задачам. Как и сорок лет назад МП предоставляет широкое поле для серьезных теоретических исследований как лингвистам, так и математикам.

СПИСОК ЛИТЕРАТУРЫ

1. Апресян Ю. Д., Богуславский И. М., Иомдин Л. Д., Крысин Л. П., Лазурский А. В., Перцов Н. В., Санников В. Э. Лингвистическое обеспечение ЭТАП-1 // Разработка формальной модели естественного языка.— Новосибирск: ВЦ СО АН, 1981.— С. 3—28.
2. Апресян Ю. Д., Цинман Л. Л. Об идеологии системы ЭТАП-II // Формальное представление лингвистической информации.— Новосибирск: ВЦ СО АН, 1982.— С. 3—19.
3. Войнов В. К., Блехман М. С. Состав, методики разработки и эффективность действующих систем машинного перевода: Учебное пособие.— Харьков: Харьковский государственный университет, 1987.
4. Гальченко О. Н., Мирам Г. Е. Система машинного перевода СИМПАР — принципы разработки и задачи экспериментальной эксплуатации // Проблемы автоматического и экспериментально-фонетического анализа текстов: Сб. научных статей.— Минск, 1986.— С. 141—146.
5. Ильин Г. М. Алгоритм синтаксического анализа для одной модели грамматики зависимостей // НТИ.— 1966.— № 1.— С. 51—53.
6. Иорданская Л. Н. Два оператора обработки словосочетаний с сильным управлением (для автоматического синтаксического анализа).— Предварительные публикации/ИЯ АН СССР.— М., 1961.— 33 с.
7. Иорданская Л. Н. О некоторых свойствах правильной синтаксической структуры (на материале русского языка) // Вопросы языкознания.— 1963.— № 4.— С. 102—112.
8. Иорданская Л. Н. Автоматический синтаксический анализ: Межсегментный синтаксический анализ.— Новосибирск: Наука, 1967.— Кибернетика в монографиях/Под общей ред. А. А. Ляпунова и О. С. Кулагиной; 2.
9. Кулагина О. С. Некоторые вопросы преобразования деревьев зависимостей.— Препринт/ИПМ АН СССР.— М., 1969.— № 12.— 30 с.
10. Кулагина О. С. Алгоритм синтаксического анализа в системе французско-русского машинного перевода.— Препринт/ИПМ АН СССР.— М., 1970.— № 13.— 47 с.
11. Кулагина О. С., Мельчук И. А., Эрастов К. О. Об одной возможной системе машинного перевода.— Предварительные публикации/ИРЯ АН СССР. Проблемная группа по экспериментальной и прикладной лингвистике.— М., 1971.— Вып. 21.— 30 с.
12. Кулагина О. С. О системе французско-русского машинного перевода ФР-II // Проблемы кибернетики: вып. 27.— М.: Наука, 1973.— С. 33—45.
13. Кулагина О. С. Исследования по машинному переводу.— М.: Наука, 1979.— 320 с.
14. Кулагина О. С. О моделировании естественных языков.— Препринт/ИПМ АН СССР.— М., 1981.— № 138.— 24 с.
15. Кулагина О. С. Об автоматическом синтаксическом анализе русских текстов.— Препринт/ИПМ АН СССР.— М., 1987.— № 205.— 22 с.

16. Лейкина Б. М., Никитина Т. Н., Откупщикова М. И., Фитилов С. Я., Цейтин Г. С. Система автоматического перевода, разрабатываемая в группе математической лингвистики ВЦ ЛГУ // НТИ.— 1966.— № 1.— С. 40—50.
17. Ловцкий Е. Е. Автоматический синтаксический анализ в системе перевода с английского языка на русский (СПАР) // Машинный перевод и автоматизация информационных процессов.— М.: Всесоюзный центр переводов, 1975.— С. 46—64.
18. Мартемьянов Ю. С. Особенности синтеза ФЛОС и их следствие для анализа // Международный семинар по машинному переводу: Тезисы докладов.— М.: Всесоюзный центр переводов, 1983.— С. 130—132.
19. Марчук Ю. Н., Тихомиров Б. Д., Щербинин В. И. Система машинного перевода с английского языка на русский // Машинный перевод и автоматизация информационных процессов.— М.: Всесоюзный центр переводов, 1975.— С. 18—33.
20. Марчук Ю. Н., Власов А. Н. Основные принципы перевода в системе немецко-русского МП НЕРПА // Международный семинар по машинному переводу: Тезисы докладов.— М.: Всесоюзный центр переводов, 1979.— С. 110—112.
21. Мельчук И. А. Автоматический синтаксический анализ: Общие принципы. Внутрисегментный синтаксический анализ.— Новосибирск: Наука, 1964.— Кибернетика в монографиях/Под редакцией А. А. Ляпунова; 1.
22. Мельчук И. А. Порядок слов при автоматическом синтезе русского текста // НТИ.— 1965.— С. 36—44.
23. Мельчук И. А. Опыт теории лингвистических моделей «смысл — текст».— М.: Наука, 1974.
24. Машинный перевод и прикладная лингвистика: Проблемы создания системы автоматического перевода: Сб. научных трудов.— М.: Мин. высш. и ср. спец. обр., МГПИИЯ, 1987; вып. 271.
25. Синтаксический компонент в системах машинного перевода: Обзорная информация.— М.: Всесоюзный центр переводов, 1981.— Машинный перевод и автоматизация информационных процессов; серия 2; вып. 5.
26. Тихомиров Б. Д. Промышленные системы машинного перевода // Научно-технический перевод.— М.: Наука, 1987.— С. 92—105.
27. Убин И. И. Перевод с помощью ЭВМ: Обзорная информация.— М.: Всесоюзный центр переводов, 1986.— Машинный перевод и автоматизация информационных процессов; серия 2; вып. 11.
28. Цейтин Г. С. Методы синтаксического анализа, использующие предпочтения языковых конструкций (модели и эксперименты) // Международный семинар по машинному переводу: Тезисы докладов.— М.: Всесоюзный центр переводов, 1975.— С. 131—133.
29. Цинман Л. Л. Развитие логико-алгоритмического обеспечения.— Предварительные публикации/ИРЯ АН СССР. Проблемная группа по эксперим. и прикладной лингвистике.— М., 1986.— Вып. 174.
30. Шаляпина З. М. Принципы формального анализа японского текста в связи с задачей его перевода // *Problemy jezykow Azji i Afryki.*— Warszawa.— 1987. P. 111—126.
31. Alonso J. A. A model for transfer control in the METAL MT-system // *COLING Budapest.*— 1988.— V. 1.— P. 19—24.
32. ALPAC. Language and Machines. Computer in translation and linguistics: Report Automatic Language Advisory Committee: Nat. Acad. Sci., Nat. Res. Council.— Washington, 1966 (Рус. пер.: Язык и машины: Отчет наблюдательного комитета по автоматической обработке текстов: Национальная Академия Наук США: Национальный научно-исслед. совет // НТИ: серия 2.— 1968.— № 8.— С. 25—36.)
33. Amano S. The Toshiba machine translation system // *Japan Computer Quarterly.*— 1986.— № 64.— P. 32—35.
34. Apello L. A compositional approach to the translation of temporal expression in Rosetta system // *COLING 86.*— 1986.— P. 313—318.
35. Arnold D. J., Krauwer S., Rosner M., Des Tombe L., Varile G. B. The (C, A)T framework in EUROTRA. A theoretically committed notation for MT // *COLING 86.*— 1986.— P. 297—303.
36. Bachut D., Verastegui N. Software tools for the environment of a computer-aided translation system // *COLING 84.*— 1984.— P. 330—333.
37. Beaven J. L., Whitelock P. Machine translation using isomorphic UCGs // *Coling Budapest.*— 1988.— P. 32—35.
38. Bennett W., Slocum J. The LRC machine translation system // *Computational linguistics.*— 1985.— V. 11, № 2—3.— P. 111—121.
39. Biewer A., Feneysel C., Ritzke J., Stegentritt E. ASCOF — a modular multilevel system for french-german translation // *Computational linguistics.*— 1985.— V. 11, № 2—3.— P. 137—154.
40. Boitet C., Guillaume P., Quezel-Ambrunaz M. Manipulation d'arborescences et parallelisme: le systeme ROBRA // *COLING 78.*— 1978.— № 24.— 5 p.

41. Boitet C., Chatelin P., Daun Fraga P. Present and future paradigms in the automatized translation of natural language // COLING 80.— 1980.— P. 430—436.
42. Boitet C. Guillaume P., Quezel-Ambrunaz M. Implementation and conversational environment of ARIANE 78.4. An integrated system for automated translation and human revision // COLING 82.— 1982.— P. 19—28.
43. Boitet C., Gerber R. Expert systems and other new techniques in MT // COLING 84.— 1984.— P. 468—473.
44. Boitet C. Pros and cons of the pivot and transfer approach in multilingual machine translation // New directions in machine translation.— 1988.— № 4.— 15 p.
45. Boitet C., Zaharin Y. Representation trees and string-tree correspondences // COLING Budapest.— 1988.— V. 1.— P. 59—64.
46. Brinkmann K. H. Machine aids translation // META.— V. 26, № 1.— P. 67—75.
47. Chung H. S., Kunii T. L. NARA: a two way simultaneous interpretation system between Korean and Japanese // COLING 86.— 1986.— P. 325—328.
48. COLING 78: 7th International Conference on Computational Linguistics: Information abstracts.— Bergen: Univ. of Bergen, 1978.
49. COLING 80: Proceedings of the 8th International Conference on Computational Linguistics.— Tokyo, 1980.
50. COLING 82: Proceedings of the Ninth International Conference on Computational Linguistics.— Amsterdam; New York; Oxford: North-Holland, 1982.— Linguistics series; № 47.
51. COLING 84: 10th International Conference on Computational Linguistics: 22nd Annual Meeting of the Association for Computational Linguistics: Proceedings of COLING 84.— Stanford: Stanford University, 1984.
52. COLING 86: 11th International Conference on Computational Linguistics: Proceedings of COLING' 86.— Bonn: Univ. of Bonn, 1986.
53. COLING Budapest: Proceedings of the 12th International Conference on Computational Linguistics.— Budapest: John von Neumann Society for Computing Sciences, 1988.— V. 1—2.
54. Dong Zhen Dong. MT research in China.— Paper presented at the COLING Budapest.— Beijing: CSTC, 1988.— 5 p.
55. Dostert B. H. User's evaluation of machine translation. Georgetown MT system. 1963—1973.— Austin: Univ. of Texas, 1973.
56. Freigang K. H. English-German machine translation as component of the multilingual machine translation system SUSY // Международный семинар по машинному переводу: Тезисы докладов.— М.: Всесоюзный центр переводов, 1983.— С. 82—83.
57. Golan I., Lappin S., Rimon M. An active bilingual lexicon for machine translation // COLING Budapest, 1988.— V. 1.— P. 205—211.
58. Guzman de Rojas I. ATAMIRI — interlingual MT using the Aymara language // New directions in machine translation, 1988.— № 6.— 6 p.
59. Hajicova E., Kirshner Z. Fail-soft («emergency») measures in a production-oriented MT system // Third Conference of the European Chapter of the Association for Computational Linguistics: Proceedings of the Conference.— Copenhagen: Univ. of Copenhagen, 1987.— P. 104—108.
60. Hajic J. RUSLAN — an MT system between closely related languages // Third Conference of the European Chapter of the Association for Computational Linguistics: Proceedings of the Conference.— Copenhagen: Univ. of Copenhagen, 1987.— P. 113—117.
61. Harada T. NEC's machine translation system «PIVOT» // Japan Computer Quarterly.— 1986, № 64.— P. 24—31.
62. Hippe Z., Kaczmarec A., Ampel T., Pablikowska B. Raport z badan nad komputer systemem do tlumaczenia textow chemicznych z jezyka rosjskiego na jezyk polski.— Rzeszow: wid. P. Rz., 1987.
63. Hornby A. S. Guide to patterns and usage in English.— London: Oxford univ Press, 1975.
64. Hutchins W. J. The evolution of machine translation systems // Practical experience of machine translation/Ed. by V. Lawson.— North-Holland, 1982.— P. 21—38.
65. Hutchins W. J. Recent development in Machine Translation. A review of the last five years // New directions in machine translation, 1988.— № 1.— 56 p.
66. Iordanskaya L. N., Kittredge R., Polguere A. Implementing a Meaning-Text Model for language generation.— Paper presented to the 12th International Conference of Computational Linguistics (COLING 88), 1988.— 5 p.
67. Isabelle P., Bourbeau L. TAUM-AVIATION: its technical features and some experimental results // Computational Linguistics.— 1985.— V. 11, № 1.— P. 18—27.
68. Isabelle P., Macklovitch E. Transfer and MT modularity // COLING 86, 1986.— P. 115—117.
69. Isabelle P., Dymetman M., Macklovitch E. CRITTER: a translation system for agricultural market reports // COLING Budapest, 1988, P. 261—266.

70. Jin W., Simmons R. F. Symmetric rules for translation of English and Chinese // *Computers and Translation*.— V. 1, № 3.— P. 153—167.
71. Johnson R. L., Krawter S., Rosner M. A., Varile G. B. The design of the kernel architecture of the EUROTRA system // *COLIGN 84, 1984*.— P. 226—235.
72. Johnson R., King M., Des Tombe L. EUROTRA: a multilingual system under development // *Computational Linguistics*.— 1985.— V. 11, № 2—3.— P. 155—169.
73. Kakigahara K., Aizawa I. Completion of Japanese sentences by inferring function words // *COLING Budapest, 1988*.— P. 291—296.
74. King M. EUROTRA: an attempt to achieve multilingual MT // *Practical experience of Machine Translation/Ed. by V. Lawson*.— North-Holland, 1982.— P. 139—148.
75. Kirshner Z. On a device in dictionary operations in machine translation // *COLING 82, 1982*.— P. 157—166.
76. Kittredge R., Polguere A., Goldberg E. Synthesizing weather forecasts from formatted data // *COLING 86, 1986*.— P. 563—565.
77. Krollman F. Computer aids to translation // *META*.— 1981.— V. 26, № 1.— P. 85—94.
78. Kunze J. Teoretische Probleme der automatischen Übersetzung // *Zeit. math. Logic und Grundl. Math.*— 1966.— V. 12, № 1—2.— P. 85—130.
79. Kunze J. Temporal relations in texts and time logical inferences // *COLING 86, 1986*.— P. 350—352.
80. Kushima H., Ushida H. Fujitsu's machine translation systems // *Japan Computer Quarterly*.— 1986.— № 64.— P. 19—32.
81. Lau P., Perschke S. Morphology in the EUROTRA base level concept // *Third Conference of the European Chapter of the Association for Computational Linguistics: Proceedings of the Conference*.— Copenhagen: Univ. of Copenhagen, 1987.— P. 19—25.
82. *Practical experience of Machine Translation/Ed. by V. Lawson*.— North-Holland, 1982.— 199 p.
83. Lecerf Y. L'adressage intrinseque en traduction automatique // *Traduction automatique*.— 1964.— V. 2, № 2—3.— P. 31—47.
84. McGee Wood M., Chandler B. J. Machine translation for monolingual // *COLING Budapest, 1988*.— P. 760—763.
85. McWilliams J. Machine translation systems at EXPO'85 // *Japan Computer Quarterly*.— 1986.— N 64.— P. 55—58.
86. Melby A. K. Design and implementation of a computer-assisted translation system // *COLING 78, 1978*.— № 56.— 5 p.
87. Melby A. K., Smith M. R., Peterson J. ITS: Interactive translation system // *COLING 80, 1980*.— P. 424—429.
88. Melby A. K. Translators and machines — can they cooperate? // *META*.— 1981.— V. 26, № 1.— P. 23—34.
89. Melby A. K. Multi-level translation aids in a distributed system // *COLING 82, 1982*.— P. 215—220.
90. Meya M., Vidal J. An integrated model for the treatment of time in machine translation systems // *COLING Budapest, 1988*.— P. 437—441.
91. Nagao M., Tsujii P., Yada K., Kakimoto T. An English-Japanese machine translation system of the titles of scientific and engineering papers // *COLING 82, 1982*.— P. 245—252.
92. Nagao M., Nishida T., Tsujii J. Dealing with incompleteness of linguistic knowledge in language translation — transfer and generation stage of Mu machine translation project // *COLING 84, 1984*.— P. 420—427.
93. Nagao M., Tsujii J., Nakamura J. The Japanese Government project for machine translation // *Computational Linguistics*.— 1982.— V. 11, № 2—3. P. 91—110.
94. Nagao M., Tsujii J. The transfer phase of the Mu machine translation system // *COLING 86, 1986*.— P. 97—103.
95. Nakamura J., Tsujii J., Nagao O. Grammar writing system (GRADE) of Mu-machine translation project and its characteristics // *COLING 84, 1984*.— P. 338—343.
96. Nakamura J., Nagao M. Extraction of semantic information from ordinary English dictionary and its evaluation // *COLING Budapest, 1988*.— P. 459—464.
97. *New directions in machine translation: Conference proceedings: Budapest 18—19 August 1988/D. Maxwell, K. Shubert, A. P. Witkam (eds.)*.— Budapest: John von Neumann Society for Computing Sciences, Dordrecht/Providence: Foris Publishers.— 1988.
98. Nishida F., Takamatsu S., Kuroki H. English — Japanese translation through case-structure conversion // *COLING 80, 1980*.— P. 447—452.
99. Nishida F., Fujita Y., Takamatsu S. Construction of a modular and portable translation system // *COLING 86, 1986*.— P. 649—651.
100. Nitta Y., Okajima A., Yamano F., Ishihara K. A heuristic approach to English-into-Japanese machine translation // *COLING 82, 1982*.— P. 283—288.

101. Nogami H., Yoshimura Y., Amano S. Parsing with look-ahead in real-time on-line translation system // COLING Budapest, 1988.— V. 2.— P. 488—493.
102. Nomura H., Naito S., Katagiri Y., Shimazu A. Translation by understanding: a machine translation system LUTE // COLING 86, 1986.— P. 621—626.
103. Oubine I. I., Tikhomirov B. D. The state of art in machine translation in the USSR // New directions in machine translation, 1988.— № 3.— 9 p.
104. Papp F. Empirical data and automatic analysis // COLING 82, 1982, P. 301—306.
105. Pericliev V. Handling syntactical ambiguity in machine translation // COLING 84, 1984.— P. 521—524.
106. Proszeky G. Hungarian — a special challenge to machine translation // New directions in machine translation, 1988.— № 14.— 13 p.
107. Root R. A two-way approach to structural transfer system in MT // Second Conference of the European Chapter of the Association for Computational Linguistics: Proceeding of the Conference.— Geneva: Univ. of Geneva, 1985.— P. 70—72.
108. Rosner D. When Mariko talks to Siegfried — experiences from a Japanese-German machine translation project // COLING 86, 1986.— P. 652—654.
109. Rothkegel A. Pragmatics in machine translation // COLING 86, 1986.— P. 335—337.
110. Ruffino J. R. Coping with machine translation // Practical experience of machine translation/Ed. by V. Lawson.— North-Holland, 1982.— P. 57—60.
111. Sawai S., Fukushima H., Sugimoto M., Ukai N. Knowledge representation and machine translation // COLING 82, 1982.— P. 351—356.
112. Schenk A. Idioms in the Rosetta machine translation system // COLING 86, 1986.— P. 319—324.
113. Sereda S. P. Practical experience of machine translation // Practical experience of machine translation/Ed. by V. Lawson.— North-Holland, 1982.— P. 119—126.
114. Sgal P., Panevova J. Machine translation, linguistics and interlingua // Third Conference of the European Chapter of the Association for Computational Linguistics: Proceedings of the Conference.— Copenhagen: Univ. of Copenhagen, 1987.— P. 99—103.
115. Sigurd B. Translating to and from swedish by SWETRA — a multilingual translating system // New directions in machine translation, 1988.— № 13.— 13 p.
116. Slocum J. A survey of Machine Translation: its history, current status and future prospects // Computational Linguistics.— 1985.— V. 11, № 1.— P. 1—17.
117. Slocum J., Bennett W. S., Whiffin L., Norcross E. An evaluation of METAL: the LRC machine translation system // Second European Chapter of the Association for Computational Linguistics: Proceedings of the Conference.— Geneva: Univ. of Geneva, 1985.— P. 62—69.
118. Somers H. MT Summit, Hakone, Kanagava-ken, Japan, September 18—19, 1987 // Computers and Translation.— 1988.— V. 3, № 1.— P. 91—95.
119. Thoun B. The METEO system // Practical experience of Machine Translation/Ed. by V. Lawson.— North — Holland, 1982.— P. 39—44.
120. Tomita M., Carbonell J. S. Another stride towards knowledgebased machine translation // COLING 86, 1986.— P. 633—642.
121. Tong L. C. English-Malay Translation system: a laboratory pototype // COLING 86, 1986.— P. 639—642.
122. Tsujii J., Nakamura J., Nagao M. Analysis grammar of Japanese in the Mu-project — a procedural approach to analysis grammar // COLING 84, 1984.— P. 267—274.
123. Tsujii J., Muto Y., Ikeda Y., Nagao M. How to get preferred reading in natural language analysis // COLING Budapest, 1988.— P. 683—687.
124. Tsujii J., Nagao M. Dialogue translation vs. text translation // COLING Budapest, 1988.— P. 688—693.
125. Tsutsumi T. A prototype English-Japanese machine translation system for translation IBM computer manuals // COLING 86, 1986.— P. 646—648.
126. Ushida H., Sugiano K. A machine translation system from Japanese into English based on conceptual structure // COLING 80, 1980.— P. 455—462.
127. Van Eynde F. The analysis of tense and aspect in EUROTRA // COLING Budapest, 1988.— P. 699—704.
128. Van Munster E. An integrated model for the treatment of time in machine translation systems // COLING Budapest, 1988.— P. 437—441.
129. Van Slype G. Economic aspects of machine translation // Practical experience of machine translation.— North — Holland, 1982.— P. 79—94.
130. Vasconcellos M., Leon M. SPANAM and ENGSPAN: machine translation at the Pan American Health Organization // Computational Linguistics.— 1985.— V. 11, № 2—3.— P. 122—136.
131. Vauquois B. La traduction automatique à Grenoble.— Paris: Dunod, 1975.— Documents de linguistique quantitative: N 24.
132. Vauquois B. L'évolution des logiciels et des modeles linguistiques pour la traduction automatisée.— Grenoble: Groupe d'étude pour la traduction automatique, 1977. (Рус. пер.: Вокуча Б. Развитие формального аппарата и лингвистических

- моделей в системах автоматического перевода // Актуальные вопросы практической реализации системы автоматического перевода.— М.: Изд-во МГУ.— 1982.— Ч. 1.— С. 24—52).
133. Vauquois B., Boitet C. Automated translation at Grenoble university // Computational Linguistics.— 1985.— V. 11, № 1.— P. 28—36.
 134. Vauquois B., Chappuy S. Static grammars: a formalism for the description of linguistic models // Proceedings of the Conference on theoretical and methodological issues in machine translation of natural language.— Hamilton: Colgate univ., 1985.— P. 298—322.
 135. Wilks Y., Fass D., Guo C., McDonald J. E., Plate T., Slator B. Machine tractable dictionaries as tools and resources for natural language processing // COLING Budapest, 1988.— V. 2.— P. 750—755.
 136. Witkam T. DLT—an industrial R&D project for multilingual translation // COLING Budapest, 1988.— P. 756—759.
 137. Woods W. Transitional network grammars for natural language analysis // Comm. ACM.— 1970.— V. 13, № 10.— P. 591—606. (Рус. пер.: Вудс В. Сетевые грамматики для анализа естественных языков // Кибернетический сборник. Новая серия.— М.: Мир, 1976.— Вып. 13.— С. 120—158).
 138. Yngve V. A framework for syntactic translation // Mechanical translation.— 1957.— V. 4, № 3.— P. 59—65.
 139. Yoshida S. A consideration on the concepts structure and language — in relation to selections of translation equivalents of verbs in machine translation systems // COLING 84, 1984.— P. 167—169.
 140. Zajac R. Interactive translation: a new approach // COLING Budapest, 1988.— P. 785—790.