

© 1996 г. З.М. ШАЛЯПИНА

АВТОМАТИЧЕСКИЙ ПЕРЕВОД: ЭВОЛЮЦИЯ И СОВРЕМЕННЫЕ ТЕНДЕНЦИИ*

Идея автоматического перевода (АП) высказывалась, как известно, еще Лейбницем – более 300 лет тому назад. Однако на реальную научную и техническую основу эта задача была поставлена только в нашем веке.

Всю первую половину XX века идея перевода с помощью искусственных механизмов "носила в воздухе". Согласно [СМП 1962], в 1924 г. в эстонской газете "Vaba maa" было опубликовано сообщение о демонстрации модели пишущей машины-переводчика изобретателем А. Вахером. В 1933 г. развернутый проект "переводческой машины" предложил российский изобретатель П.П. Петров-Троянский – наиболее известный из предшественников современного АП [Панов 1959]. В том же 1933 г. во Франции инженер Георгий Арцруни получил патент на машину для перевода, которую он назвал "Механическим мозгом" [Corbé 1960].

Но по-настоящему серьезная работа в этой области началась только в конце 40-х – начале 50-х годов, после появления электронно-вычислительной техники.

Принципы АП, господствовавшие в тот период, с современных позиций могли бы быть охарактеризованы как обработка текста под управлением словаря (lexicon-driven processing), которая осуществляется с помощью однозначных процедур, целиком подчинена задаче перехода от входного языка к выходному (т.е. не предусматривает никаких этапов, чьи функции были бы сколько-нибудь самостоятельными и могли бы определяться не в связи с этим межъязыковым переходом) и не использует эксплицитно выделенных лингвистических правил, но полностью инкорпорирует их в программные процедуры (так что из всей лингвистической информации явным образом задаются только перечни внутренне не структурированных признаков, приписываемые словам в словаре и определяемые лишь косвенно – через ссылки на эти признаки в использующих их процедурных проверках).

Задача АП ставилась тогда как чисто практическая. Ожидалось, что хотя бы для научно-технических текстов будут быстро построены машинные системы, обеспечивающие их перевод в промышленных масштабах. Ожидания эти, однако, не могли оправдаться, так как были основаны на двух принципиальных заблуждениях: с одной стороны, переоценивалась степень близости тех операций, в терминах которых определялось функционирование ЭВМ, к элементарным функциям человеческого интеллекта; с другой стороны – недооценивалась фундаментальная сложность естественного языка как инструмента интеллектуальной деятельности человека и перевода как одного из отнюдь не самых простых видов этой деятельности.

Первые десять лет развития данной области дали достаточно оснований для осознания этих заблуждений в качестве таковых. Однако реакция на это осознание

* Данная работа ни в коей мере не претендует на сколько-нибудь полный обзор рассматриваемой области: эта область настолько обширна и настолько интенсивно развивается, что полный обзор соответствующих работ уже за первые 15 лет ее существования [Мельчук–Равич 1967] имел объем более 32 а.л. В настоящее же время неполноты не могут избежать даже такие подробные ее рассмотрения, как [Hutchins 1986] или [Nagao 1991]. Мы должны также принести читателю свои извинения за неизбежную субъективность и схематичность изложения и за многочисленные лакуны в приводимой библиографии.

оказалось далеко не однозначной. Можно говорить о трех ее типах: 1) разочарование в самой идее автоматизации перевода; 2) поиск таких областей применения перевода, где разрыв между возможностями ЭВМ и потребностями пользователя минимален; 3) работа по преодолению этого разрыва.

Первый тип реакции – разочарование в идее АП – был характерен прежде всего для представителей научно-технической администрации. Наиболее явно он был выражен в докладе, представленном в 1966 г. Конгрессу США Консультативной комиссией по автоматической обработке естественного языка Национальной академии наук [ALPAC 1966]. Выводы комиссии сводились к тому, что по меньшей мере в ближайшие 50 лет разработки по машинному переводу не могут дать какого-либо практического выхода, а основанные на них системы останутся принципиально нерентабельными. В последующие 10–15 лет научно-техническая политика в области АП почти повсеместно строилась на базе этих выводов. В Европе и СССР работы по созданию практических систем АП были практически свернуты, в самих США они продолжались, но без прежнего энтузиазма, в основном в рамках лишь тех разработок, которые были начаты до опубликования доклада.

Для разработчиков, непосредственно создававших практические системы АП, более типичным явился второй тип реакции: поиск тех сфер применения таких систем, в пределах которых они имитируют поведение человека достаточно адекватно и дают результаты, его удовлетворяющие.

Наиболее очевидной из этих сфер явился словарный поиск и те виды перевода, где он играет определяющую роль: начиная с "черновых" переводов, используемых для предварительного просмотра текста и определения его тематической принадлежности, а также степени его релевантности для нужд пользователя (с тем, например, чтобы решить, стоит ли заказывать перевод этого текста переводчику-человеку), и кончая переводами таких специфических текстов, как, например, анкетные данные или инвентарные списки. Так, согласно [Слокум 1989], в Комиссии по атомной энергии США и в Евратоме для получения "черновых" переводов на английский язык работ русских исследователей, уже с 1964 г. стала применяться система GAT, а в Отделе зарубежных технологий ВВС США – разработанная фирмой IBM система Mark-II; в 70-х гг. обе системы были заменены созданной на базе GAT системой SYSTRAN.

Для более общих переводческих задач использование машинных систем также оказалось продуктивным – хотя, возможно, в основном за счет автоматизации различных вспомогательных операций: прежде всего опять-таки операций словарного поиска и различных процедур постредактирования текста перевода, но кроме того – по мере их описания и отладки – также стандартных механизмов морфологической и синтаксической обработки входного и выходного текстов. Все это ускоряет работу переводчика-человека и тем самым удешевляет весь процесс перевода. Так, система SPANAM, применяющая целый ряд вспомогательных средств и механизмов, облегчающих постредактирование, позволила ускорить перевод, в зависимости от текста, в 1,5–4 раза [Vasconcellos-León 1985]. Англо-французская и англо-испанская версия системы SYSTRAN, используемые корпорацией "Дженерал Моторс", привели к повышению производительности труда переводчиков в 3–4 раза [Слокум 1989]. Японско-английская система перевода аннотаций научных работ, созданная под руководством М. Нагао в рамках Японского национального проекта по АП (так наз. MU-проекта [Nagao et al. 1985]) и работающая в Японском информационном центре по науке и технике, по сведениям, приводимым в [Кау et al. 1994], сократила затраты на перевод и его редактирование в два раза.

Помимо систем АП как таковых, в целях автоматизации отдельных переводческих операций стали создаваться также более частные "машинные средства" в помощь переводчику и редактору: автоматические словари и терминологические базы данных, компьютерные тезаурусы, средства экранного редактирования, системы орфографической, терминологической и грамматической коррекции текстов и т.п.

С другой стороны, обнаружилось, что для текстов, характеризующихся относительно жесткой структурой, настроенные на них системы АП могут давать вполне приемлемые переводы и без постредактирования, а в отдельных случаях даже вообще без вмешательства человека. Простейшим видом систем такого рода являются автоматические разговорники для туристов, предлагающие пользователю более или менее разнообразные "меню" стандартных вопросов и ответов на двух или нескольких языках. Из более сложно организованных систем АП для жестко структурированных текстов наиболее известна система TAUM-METEO, переводящая прогнозы погоды с английского на французский язык в Канадском метеорологическом центре. К настоящему времени, согласно [Kay et al. 1994: 41], построена расширенная версия этой системы, на вход которой подаются уже не тексты прогнозов, а только релевантные для них показания метеорологических приборов. Система автоматически составляет по этим данным и английские и французские варианты требуемых прогнозов – т.е. "переводит" уже не с одного естественного языка на другой, а с языка "формальной семантики" данной области на естественные языки.

То, что для текстов с жесткими структурными ограничениями АП оказался более успешным, чем для текстов произвольной структуры, послужило стимулом для формирования и развития еще одной сферы промышленного и коммерческого применения систем АП – применения их для перевода с пред- и/или интерредактированием входного текста в соответствии с требованиями системы. Из различных способов такого пред- и интерредактирования наиболее удачным оказался тот, при котором пользователь предварительно или постепенно – с учетом вопросов или рекомендаций, поступающих по ходу работы от самой системы АП, – (пере)формулирует переводимый текст в терминах того специализированного варианта входного языка, в пределах которого система обеспечивает достаточно надежный выход. Такой подход дает хорошие результаты при переводе с родного языка пользователя и особенно привлекателен тем, что знание выходного языка для него при этом необязательно.

Известно, что фирма XEROX успешно эксплуатирует в таком режиме систему SYSTRAN: система переводит внутреннюю техническую документацию фирмы, которую требуется составлять в этих целях на ограниченном варианте английского языка – так называемом Multinational Customized English, – имеющем фиксированный словарь и строго определенный набор синтаксических структур. Есть сведения об аналогичном применении ограничений на входной язык фирмами Caterpillar и Boeing. В [Nagao 1991] описан эксперимент по использованию подобным способом также японско-английской системы ATLAS II. К настоящему времени начали разрабатываться и специальные системы АП для работы с пользователем, знающим один только входной язык (monolingual user): LIDIA [Blanchon 1994], N_{tran} [Whitelock et al. 1986] и др.

Выявление таких оптимальных для АП сфер его применения, построение систем, ориентированных именно на эти сферы или адаптация к ним систем, ставивших ранее другие задачи (например, переориентация разработок, начатых в университете Бригема Янга в целях перевода религиозных текстов мормонской церкви, при их продолжении в рамках корпорации Weidner Communications и фирмы ALPS, на создание интерактивных "машинных средств в помощь переводчику") и особенно накопление больших массивов словарей и грамматических правил, представленных в требуемом этими системами виде (так, словари системы SYSTRAN в Отделе зарубежных технологий ВВС США насчитывают на данный момент более миллиона единиц [Слокум 1989]), привело к концу 70-х – началу 80-х гг. к новой вспышке энтузиазма по поводу промышленной и коммерческой ценности АП.

Среди прочих факторов этому способствовали, с одной стороны, такие яркие свидетельства применимости АП для решения реальных переводческих задач, как использование системы SYSTRAN для перевода на английский язык русской космической документации в рамках программы советско-американского сотрудничества "Аполлон-Союз" (1974 г.), а с другой – активное включение в разработку систем АП японских компьютерных и переводческих фирм, которые уверенно

осваивают данную область и все более успешно конкурируют с ведущими фирмами-разработчиками таких систем в США и Европе (достаточно упомянуть, что в 1983 г. японская компания Bravice приобрела контрольный пакет акций, фирмы Weidner Communications, а в настоящее время японские вкладчики владеют и значительной частью акций фирмы SYSTRAN). Однако в собственно научном плане развитие коммерческих систем АП было относительно мало результативным: как отмечается в [Kay et al. 1994], наибольший коммерческий успех в 80-х гг. пришелся на долю системы SYSTRAN – при том, что научные принципы, положенные в основу ее разработки в 60-х гг., с тех пор практически не модернизировались.

Для научного развития наибольшее значение имела, как и следовало ожидать, реакция третьего типа – понимание, что имеющихся научных знаний недостаточно для построения систем, гарантирующих стабильно высокое качество перевода, и отношение к этому как к научному вызову, ставящему перед исследователями задачу анализа и преодоления обнаруженного разрыва между возможностями ЭВМ и спецификой естественного языка и языковой деятельности человека.

Работа по решению данной задачи стала вестись с двух сторон.

С одной стороны, начались исследования в области создания логико-математических и программных средств высокого уровня, которые позволяли бы задавать машинные процедуры и используемую ими информацию в формате, более прозрачном, компактном и удобном для человеческого восприятия – с тем, чтобы разработчик мог сосредоточиться на содержании решаемой задачи и не отвлекаться на технические проблемы записи своего решения.

С другой стороны, стали разрабатываться формальные модели естественного языка и языковой деятельности, в которых языковые факты и явления стали осмысляться с точки зрения их возможной машинной обработки и представляться в форме, удобной для целей такой обработки. И, поскольку эта обработка ведется в терминах очень простых типов операций, а языковые явления, напротив, очень сложны, многослойны, синкретичны, – то лингвистика тем самым получила мощный толчок к выходу на новый уровень научного анализа – расчленения своего комплексного, многомерного, синкретического объекта на более простые, в идеале одномерные компоненты и аспекты.

Парадоксальным результатом такого развития явилось то, что сам АП в рамках этих исследований ушел на второй план и стал рассматриваться скорее как одно из их возможных приложений, служить полигоном для экспериментальной проверки и верификации создаваемых логико-математических средств и формально-лингвистических моделей в общем ряду других подобных приложений: информационного поиска, компьютеризованного обучения, экспертных систем и т.п.

Но в то же время АП занял в этом ряду и свое самостоятельное место в качестве области моделирования того особого, отличного от других вида языковой деятельности человека, каким является перевод.

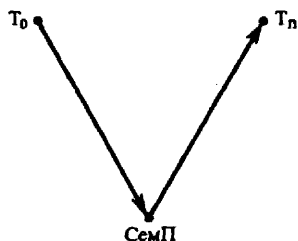
Говоря об эволюции АП в качестве такой самостоятельной области исследований, нужно прежде всего отметить, что эволюционировала и продолжает эволюционировать уже сама общая схема процесса перевода, как он предстает в системах АП.

В первых системах реализовался так называемый "прямой" подход к переводу, в рамках которого, как уже отмечалось, все осуществляемые при переводе операции трактовались как операции межъязыкового перехода – преобразования текста оригинала T_0 в текст перевода T_n .

$$T_0 \bullet \text{=====} \bullet T_n$$

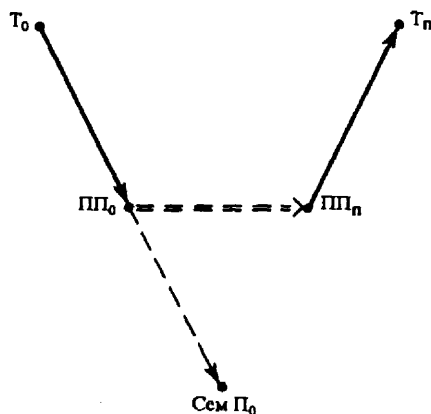
Но уже в конце 50-х гг. возникла идея перевода через особое представление текста – интерлингву (Richens 1956), или язык-посредник (ЯП) [Андреев 1957]. ЯП мыслился вначале как некоторое универсальное представление, равно удаленное от всех естественных языков, – представление, которое позже в одной из его интер-

претаций стали называть смысловой записью [Жолковский–Леонтьев–Мартемьянов 1961], или семантическим представлением СемП [Мельчук 1974]. Процесс перевода разбивался тем самым на два этапа: анализ, обеспечивающий переход от текста оригинала к его СемП, и синтез, порождающий по этому СемП текст перевода:



Достаточно скоро стало очевидно, однако, что и такая схема перевода неоптимальна. С одной стороны, создать универсальный ЯП нереально, с другой – человек явно действует не этим методом: иначе не возникало бы при переводе обычных для него смысловых сдвигов, а также ошибок, вызванных влиянием на построение текста перевода средств выражения, имевшихся в тексте оригинала. Логично постулировать, следовательно, некоторый промежуточный уровень представления и того и другого текста (resp., $ПП_0$ и $ПП_n$), – где входной и выходной языки еще не совпадают, но в то же время их соотношение уже более прозрачно, чем в плане выражения. На этом уровне (выбор которого допускает, вероятно, варьирование в зависимости от конкретной переводческой ситуации) и должен выполняться собственно перевод – операции перехода от единиц входного текста к их переводным эквивалентам.

По данной схеме, система вначале осуществляет анализ текста оригинала в целях построения его представления требуемого уровня. При этом она может в принципе обращаться и к семантическому представлению $СемП_0$ этого текста – т.е. к сведениям о его плане содержания, – но не эксплицировать полностью это содержание, а сразу по получении нужного промежуточного представления выполнять межъязыковой переход. Он дает, в свою очередь, промежуточное представление, но уже выходного текста, по которому и производится далее синтез самого этого выходного текста [Шаляпина 1975; Vauquois 1976]:



Модель стала, таким образом, принципиально трехчленной:

анализ → межъязыковые операции → синтез.

С середины 70-х гг. эта трехчленная модель получила всеобщее признание, практически вытеснив обе предыдущие, и дальнейшее развитие шло уже по линии ее

уточнения и усложнения. Понятие языка-посредника продолжает использоваться в тех из современных работ по АП, которые претендуют на достаточно глубокое проникновение в содержание переводимого текста. Однако этап межъязыкового перехода при этом, как правило, не исключается – так что данный термин является по существу просто обозначением одного из вариантов того промежуточного уровня структурного представления текста, на котором выполняется такой переход.

Определение уровня межъязыкового перехода в разных системах АП может варьировать как по признаку соотношения этого уровня с уровнями выражения/содержания языковых единиц, так и по признаку размерности тех единиц, которые обрабатываются в его рамках.

С точки зрения соотношения плана выражения и плана содержания, уровень межъязыкового перехода – это, как правило, такое представление структуры переводимых единиц, в котором сами структуры – т.е. характер и типы структурных отношений – являются достаточно абстрактными и в этом смысле более или менее универсальными.

Интересно, что структуры составляющих, которым в системах АП долгое время отдавалось в этом качестве явное предпочтение, все более вытесняются теперь другими видами структур и грамматик.

Очень распространены сетевые структуры – в том числе опирающиеся на формализм фреймов (напр., в системах SEMSYN [Rösner 1986], LUTE [Nomura et al. 1986], KANT [Nyberg–Mitamura 1992], Shalt2 [Takeda et al. 1992], на деревья зависимостей (SUSY [Luckhart–Maas 1983], GETA [Vauquois–Boitet 1985], ATHENE (в коммерческой версии – HICATS) [Кадзи–Окадзима 1985], MU [Nagao et al. 1985], ЭТАП-2 [Апресян и др. 1989] и др.), а также на различные варианты совмещения зависимостей и составляющих – обобщенные или расширенные грамматики составляющих ([Whitelock et al. 1986], [Choi et al., 1994], [Morimoto et al. 1992]), унификационные и лексико-функциональные грамматики ([Kudo–Nomura 1986], [Sharp 1989], [Noord van 1990], [Beaven 1992], [Cornish et al. 1994]), грамматики, формулируемые в терминах расширенных сетей переходов [Amano 1986], и т.п. Экспериментируют также с представлениями, использующими логику предикатов в сочетании с падежной грамматикой [Nishida et al. 1986], грамматики Монтегю (система Rosetta [Appelo 1986]), структуры, совмещающие грамматики Монтегю с составляющими и зависимостями [Gunji 1987].

Типы отношений, используемые в таких структурах, – т.е. метки на дугах (ветвях) – могут отражать поверхностный синтаксис языка – как, например, в системе SHALT [Tsutsumi 1986]). Но чаще используется классификация синтагматических отношений по смысловым основаниям: "глубинные падежи" (как в MU-проекте), "понятийные" или "концептуальные" зависимости (как в системах HICATS или MOPTRANS [Hutchins 1986]) и т.п.

Как раз о таких структурах часто и говорят как о "смысловом представлении" текста [Nomura et al. 1986] или о языке-посреднике [Harada 1986]).

Однако узлы в таких структурах – их "лексика", – в отличие от классического понимания семантического представления или языка-посредника, практически всегда привязаны к конкретному естественному языку. Чаще всего это просто лексемы соответствующего языка, но с уточнением, в каком из их значений они взяты. Нередко они называются при этом "понятиями" или "концептами" (как, например, в системе Thalia-3 [Окаси и др. 1986]). Иногда, впрочем, используется лексика, отличная и от входного и от выходного языка. Так, в системе DLT [Paragaaij 1986], в этом качестве взят модифицированный эсперанто, а в японско-немецкой системе SEMSYN [Rösner 1986] промежуточное представление опирается на английскую лексику. Ясно, однако, что "универсальность" такого ЯП – скорее вопрос удобства, чем его принципиальное свойство, по которому он противостоял бы структурным представлениям на естественном языке. В частности, в системе SEMSYN английская лексика выбрана для него просто потому, что анализатор японского языка заимствован в этой системе из японско-английской системы ATLAS-2 [Утида 1985].

Еще один вариант определения лексики ЯП – это трактовка в качестве таковой соответствий между единицами входного (входных) и выходного (выходных) языков (система TRANSLATOR [Nirenburg et al. 1986]). Здесь тоже очевидна привязка к естественным языкам – так что к семантической записи как к представлению плана содержания лексических единиц полностью не переходит ни одна система АП. В то же время все они используют более или менее сложные семантические характеристики этих единиц: от простейших перечней семантических дескрипторов до формул грамматики Монтегю [Landsbergen 1987], фреймов и сценариев [Tomita–Carbonell 1986], семантических пред- и пост-условий [Cornish et al. 1994] и т.п.

Сказанное характеризует уровень межъязыковых операций с точки зрения его места на оси "выражение–содержание" языковых единиц.

Что же касается объема обрабатываемых на этом уровне единиц – т.е. их места в иерархии синтагматического членения, – то здесь вариаций намного меньше.

Как правило, обработка ограничивается уровнем предложения. За его пределы выходят лишь некоторые чисто экспериментальные разработки, но и они, как правило, не предусматривают построения связных представлений сверхфразовых единиц, а только накапливают и хранят информацию о единицах, которые могут понадобиться для учета анафоры, анализа категории определенности (в целях правильной расстановки артиклей), корректировки порядка слов в переводе и т.п. Примером может служить система LUTE [Nomura et al. 1986], а также система фирмы Мацусита, описанная в [Cornish et al. 1994] (последняя, в частности, следит за "текущим" и "потенциальными" фокусами текста). Теоретические и экспериментальные разработки в этом направлении ведутся в Карловом университете в Чехии [Hajičová–Vrbová 1982], в Манчестерском университете в Великобритании [Whitelock 1992] и в других исследовательских центрах.

Но модель перевода развивается не только в части уточнения и расширения тех лингвистических знаний о языке и тексте, с использованием которых осуществляются собственно переводческие операции. Пополняется и усложняется и сам комплекс включаемых в нее процедурных компонентов.

Начать с того, что в этой модели появилась уже и правая "ножка" – частичное обращение к семантике в рамках работы не с входным, но с выходным языком.

Правда, до сих пор работа с семантикой выходного языка осуществляется в системах АП не наряду с учетом семантики входного текста, а вместо него. Так работает, например, англо-японская система ATLAS-1 [Kushima et al. 1986]: в ней вся используемая семантическая информация приписана не входным английским единицам, но только их японским эквивалентам. Аналогичный подход прорабатывается в группе Р.Г. Пиотровского для многоязычной системы СИЛОД [Беляева 1989], в рамках которой предлагалось осуществлять семантическую обработку текстов через Многоцелевой Автоматический Русский Словарь MAPC, независимо от того, является русский язык входным или выходным.

Такой подход оправдан гипотезой об универсальном характере семантики естественного языка, в отличие от его более поверхностных аспектов. Эта гипотеза, однако, верна лишь до определенного предела, за которым начинает действовать обратное утверждение: "... существует гораздо большее количество понятий..., которые так органически сплетены со своим языком, что... без искажения не могут быть перенесены в другие языки" [Гумбольдт 1984: 319]. Поэтому в более общем случае добавление семантического анализа в процедуры синтеза должно было бы иметь и вполне самостоятельное значение: не (только) как средство восполнения неполноты семантической информации о входном языке за счет аналогичной информации о языке перевода, но (и) как способ верификации адекватности перевода как такового – оценки того, насколько точно передают смысл оригинала те лексические и структурные эквиваленты составляющих его входных единиц, которые были выбраны в процессе межъязыковых операций. В такой самостоятельной роли, однако, семантический анализ языка перевода пока не реализован.

Другое дополнение исходной трехчленной схемы перевода, которое уже более принципиально для функционирования современных систем АП, – это включение в данную схему перифрастических компонентов. Операции перифразирования осуществляются на том же уровне, что и межъязыковой переход, и могут выполняться как до, так и после него. И по- и после-переводческое перифразирование реализовано, например, в японско-английской части МУ-проекта [Нагао и др. 1986]. Так, определительные конструкции с японским глаголом *моцу* "иметь" преобразуются перед их переводом на английский язык в конструкции с предлогом принадлежности, что позволяет вместо:

equations that have logarithmic characteristics получать в переводе:

equations with logarithmic characteristics.

Напротив, такие перифрастические преобразования, как "стяжение" каузативных конструкций, т.е. замены типа:

A makes B rotate $\Rightarrow$ A rotates B

производятся уже после перевода, в рамках английского структурного представления.

В некоторых системах применяются только после-переводческие перифрастические преобразования. Например, в системе SHALT2 [Takeda et al. 1992] они служат для трансформаций типа: *John is a good swimmer $\Rightarrow$ John swims well.*

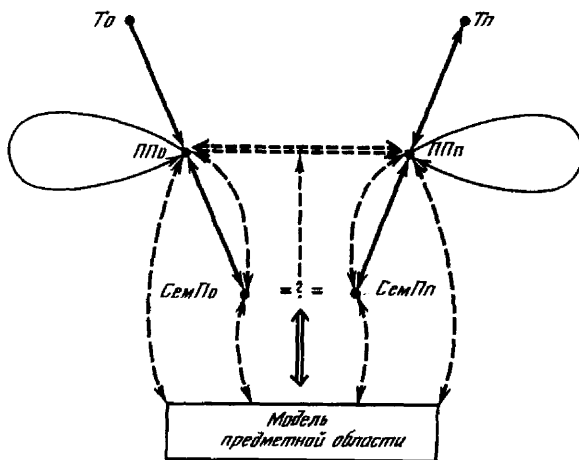
Заметим, что по сравнению с теми возможностями, которые предусматриваются моделью семантического синтеза, опирающейся на понятие лексических функций [Жолковский–Мельчук 1967], перифрастические компоненты существующих систем АП весьма ограничены и охватывают лишь малую долю перифраз, исчисляемых этой моделью. Работы, претендующие на более полную ее реализацию (типа [Neuylen 1994]), остаются еще большей частью на уровне теоретических исследований.

В то же время весьма полно реализуется третий вид перифразирования, в теоретических работах данного направления почти не рассматривавшийся: межъязыковые перифрастические операции, составляющие важный аспект межъязыковых преобразований. Во многих системах как раз этот вид перифразирования восполняет отсутствие или недостаточность до- и после-переводческих перифрастических преобразований. Такова, в частности, система ЭТАП-2 [Апресян и др. 1989].

Последний вид дополнений к общей трехчленной схеме АП в ее современных интерпретациях – это подключение к процессу перевода, помимо сведений о языке как таковом, также информации, не являющейся собственно лингвистической, – "знаний о мире". Эта информация привлекается, как правило, либо после, либо в процессе построения структурного представления входного текста при его анализе.

Системы, использующие "знания о мире", составляют, по мнению ряда авторов, отдельное направление в АП, – так называемый Knowledge-Based Machine Translation. Чаще всего они ориентированы на достаточно узкие предметные области. Например, система, описываемая в [Tomita-Carbonell 1986], моделирует перевод диалогов между врачом и пациентом, пришедшим к нему на прием. Обычно системы данного направления учитывают и лингвистические знания, причем в достаточно полном объеме, но в них часто действуют другие приоритеты. Так, по сведениям [Hutchins 1986], в системе MOPTRANS, использующей "пакетные" и "абстрагирующие" знания типа "сценариев", исходной при обработке текста является семантическая информация, а синтаксические сведения служат уже для выбора между разными семантически приемлемыми вариантами формируемой структуры (в отечественной компьютерной лингвистике похожее направление развивается в связи с созданием естественно-языковых интерфейсов к базам данных [Нариньяни 1995]).

Как можно судить уже по данному (заведомо неполному) обзору, общий набор компонентов в современных системах АП в целом приблизился к тому теоретическому их исчислению, которое обосновывалось в свое время в [Шаляпина 1980]:



Однако если посмотреть, каков удельный вес разных компонентов этой расширенной схемы перевода в современных системах АП, то приоритет по-прежнему остается за анализом. В материалах всех последних конференций по компьютерной лингвистике секция анализа содержала по крайней мере вдвое больше работ, чем каждая из остальных секций.

Любопытно заметить в этой связи, что почти за 50 лет развития АП исследования по анализу текстов описали в целом ряде отношений полный виток спирали. Это касается прежде всего общей организации процедур анализа и используемой в них лингвистической информации.

Выше отмечалось, например, что первые системы АП начинали с использования однозначных процедур анализа, строивших для обрабатываемого текста всякий раз только один вариант его представления. Затем, в 60–70-е годы интерес исследователей переключился на множественные стратегии анализа [Кулагина–Мельчук 1971]. В последнее же десятилетие вновь наблюдается переход к однозначным стратегиям, но уже на новом уровне научного осмысления проблемы. Можно отметить три основных направления в достижении такой однозначности.

Во-первых, используется так называемое "упакованное" представление неоднозначностей – т.е. неоднозначные элементы и структуры текста не "расщепляются" на разные дорожки анализа, но сохраняют формально свою "целостность" и только получают в составе формального представления анализируемого текста особые характеристики. Такой подход обосновывается, например, в [Whitelock et al. 1986; Takeda et al. 1992] и др. Среди отечественных систем весьма сходная методика используется в системе АМПАР [ИИ 1990 249–252].

Во-вторых, очень широко распространены стратегии анализа, использующие так называемую "тактику наилучших догадок" (best guess policy), т.е. учитывающие различные весовые характеристики формируемых структур (с применением статистики и разного рода эвристик) и допускающие возврат к "развилкам", возникавшим на предыдущих этапах анализа, если путь, выбранный в качестве основного, заходит в тупик. Такие стратегии разрабатываются в рамках уже упоминавшихся систем TRANSLATOR, SUSY, SHALT-2, LUTE, HICATS, системы Shake-and-bake translation [Beaven 1992] и др. Среди отечественных разработчиков, проводящих тот же подход, следует назвать прежде всего Г.С. Цейтина, которому принадлежат пионерские исследования в этом направлении [Цейтин 1975], и О.С. Кулагину [Кулагина 1992].

Наконец, третье направление исследований, ставящих задачу обеспечить однозначность анализа при АП, состоит в поиске и создании особых средств и способов организации лингвистического описания, которые позволяли бы анализатору однозначно строить правильные структуры, не прибегая ни к параллельной обработке, ни к

возвратам. В качестве одного из таких средств предложена, например, "детерминистская грамматика" [Marcus 1987].

Приоритет имеет, таким образом, не полнота охвата всех принципиально мыслимых вариантов обработки, но возможность построения нужного варианта без перебора или с минимальным перебором ложных.

Обращение к первоначальной системе приоритетов наблюдается и в области организации лингвистической информации в системах АП. В 40 – 50-е гг. лингвистическое обеспечение таких систем было фактически сплошь словарно-ориентированным. Затем произошел поворот к представлению лингвистической информации в виде грамматик, опирающихся прежде всего на правила и отношения (rule-oriented framework). В конце 60-х годов А.К. Жолковский и И.А. Мельчук вновь заявили о центральной роли словаря в описании языка [Жолковский–Мельчук 1969: 31]. Но данному утверждению суждено было остаться тогда только теоретическим постулатом. Теперь же этот постулат проникает в разработки по АП в виде объектно-ориентированных и сущностно-ориентированных подходов, лингвистическая суть которых, согласно [Whitelock 1992: 785], выражена утверждением: "The only meaningful elements of a grammar are its lexical items".

К системам, опирающимся на такие методы организации своего лингвистического обеспечения, относятся, например, LUTE, KBMT, SHALT-2, среди отечественных систем – японско-русская система ЯРАП (см. ее описания в [Modina–Shalyapina 1994; Модина–Шаляпина 1995]).

Из других компонентов модели перевода в последнее время приобретает все большую самостоятельность, выходя из младенческого периода чисто эмпирических описаний, сфера межъязыковых операций. Правда, она все еще не полностью сформировалась как отдельная научная область. На конференциях по компьютерной лингвистике для нее до сих пор не выделяется отдельной рубрики и соответствующие работы помещаются либо в общую секцию "Машинный перевод", либо в секции анализа или синтеза. При разработке систем АП тоже нередко наблюдаются попытки "инкорпорировать" межъязыковые операции в состав процедур анализа [Kushima–Uchida 1986] или синтеза [Chung–Kunii 1986; Танака–Мидзогути 1985]. Однако в последнее десятилетие здесь появилось и активно развивается собственное, весьма перспективное теоретико-методологическое направление – перевод "по аналогии" [Nagao 1984] и его варианты, прежде всего перевод "по примерам" (example-based transfer, см., например [Watanabe 94]) и перевод "по образцам" (case-based translation, исследуемый, например, в [Nomiya 1992]). Из отечественных исследований к данному направлению теснее всего примыкают оставшиеся пока не реализованными разработки Б.Р. Певзнера [Певзнер 1981].

Что касается синтеза и других видов операций с выходным текстом, то им при создании систем АП уделяется пока меньше внимания. Наиболее активно обсуждается здесь задача определения порядка слов в языках, где он не является фиксированным (см., например [Whitelock 1992]). Но, скажем, проблемы обработки лексически связанных словосочетаний выходного языка все еще решаются обычно не в рамках синтеза, но на этапе межъязыкового перехода. Например, в [Nomiya 1992] вопрос о том, что в английском языке *cow* ("корова") ближе к *horse* ("лошадь") в ситуации кормления, но к *dog* ("собака") в ситуации движения (*cow/horse feeds* vs. *dog eats*, но *cow/dog runs* vs. *horse gallops*), ставится в связи с проблематикой "перевода по образцам".

Тем не менее в последнее время и данная область становится предметом более глубокого рассмотрения. В частности, начинают изучаться ситуации ошибочного понимания текстов читателем и возможности предотвращения таких ошибок посредством перифразирования [Lenke 1994]. Можно ожидать поэтому, что уже в недалеком будущем синтез приобретет в системах АП ту активную роль, которая приходится на его долю, согласно модели "Смысл ↔ Текст".

В заключение подчеркнем следующие тенденции, характерные для современных разработок по АП.

В области создания практически ориентированных систем АП повышенное внимание уделяется сервисным средствам, дающим больше свободы действий разработчикам лингвистического обеспечения системы, а также предоставляющим пользователю все более широкие возможности не только для редактирования входных и/или выходных текстов, но и для участия в "обучении" самой системы – коррекции ее словарей и грамматик.

В области построения процедурных компонентов систем АП жестко фиксированные глобальные управляющие механизмы уступают место гибким стратегиям, допускающим варьирование порядка работы системы под управлением словаря и текста.

Наконец, в области организации используемых лингвистических описаний наблюдается все возрастающий интерес к использованию готовых (традиционных) источников естественно-языковой информации – прежде всего толковых и переводных словарей и грамматик, ориентированных на человека. для них создаются различные формальные процедуры в целях автоматического преобразования соответствующей словарной или грамматической информации в нужную форму. Ведется работа и по автоматизации самого процесса накопления необходимой для перевода лингвистической информации путем извлечения ее из параллельных текстов, пополнения в ходе самого АП (обычно в диалоге с разработчиком или пользователем системы) и т.д. (см., например [Miura et al. 1992 : Kumano-Hirakawa 1994] и др.).

С другой стороны, лингвистические описания, ориентированные на использование человеком, начинают все чаще опираться на тот или иной формальный аппарат. Наиболее ярким примером эффективности таких описаний является словарь-справочник по русскому словоизменению А.А. Зализняка [Зализняк 1977]: его текст может применяться как человеком, так и формальной системой, причем для последней не требуется при этом никаких предварительных преобразований соответствующих текстовых файлов (см., например [Kanovich-Shalyapina 1984]). Аналогичным образом, вполне формализованный характер имеет представление моделей управления английских слов в таких, например, словарях, как [LDOCE 1995] и т.д.

Таким образом, проблематика АП все более сближается в настоящее время с проблематикой языкознания в целом, и характерное для 60–70-х гг. XX века противопоставление "формальной" ("структурной") и "традиционной" лингвистики постепенно уступает место в данной области их сотрудничеству и взаимодействию**.

СПИСОК ЛИТЕРАТУРЫ

- Апресян Ю.Д. и др. 1989 – Лингвистическое обеспечение системы Этап-2. М., 1989.
Андреев Н.Д. 1957 – Машинный перевод и проблема языка-посредника // ВЯ. 1957. № 5.
Гумбольдт В. 1984 – Избранные труды по языкознанию. М., 1984.
Жолковский А.К. – Леонтьева Н.Н. – Мартынянов Ю.С. 1961 – О принципиальном использовании смысла при машинном переводе // Машинный перевод. М., 1961.
Жолковский А.К. – Мельчук И.А. 1967 – О семантическом синтезе // Проблемы кибернетики. Вып. 19. М., 1967.
Жолковский А.К. – Мельчук И.А. 1969 – К построению действующей модели языка "Смысл ↔ Текст" // Машинный перевод и прикладная лингвистика. Вып. 11. М., 1969.
Зализняк А.А. 1977 – Грамматический словарь русского языка. М., 1977. ИИ 1990 – Искусственный интеллект. Справочник. Книга 1. Системы общения и экспертные системы. М., 1990.
Кадзи Х. – Окадзима А. 1985 – Хитати ни окэру кикай хонгяку сисутэму. Дзехо сери. 1985. Т. 26. № 10 (на япон. яз.).
Кулагина О.С. 1992 – Синтаксический анализ на основе предпочтений // Festshrift V.Yu. Rosencveig. WSIA. Sbd. 33. 1992.
Кулагина О.С. – Мельчук И.А. 1971 – Автоматический перевод: краткая история, современное состояние, возможные перспективы // Автоматический перевод. М., 1971.
Мельчук И.А. 1974 – Опыт теории лингвистических моделей "Смысл ↔ Текст". М., 1974.

** Исследование выполнялось при поддержке Российского фонда фундаментальных исследований и Российского гуманитарного научного фонда.

- Мельчук И.А. – Равич Р.Д. 1967 – Автоматический перевод. 1949–1963. Аналитико-библиографический справочник. М., 1967.
- Модина Л.С. – Шалапина З.М. 1995 – Принципы организации лингвистических знаний в объектно-ориентированной модели лексико-морфологической системы японского языка // DIALOG'95. Труды Международного семинара по компьютерной лингвистике и ее приложениям. Казань, 1995.
- Нагао М. и др. 1986 – Машинный перевод с японского языка на английский // ТИИЭР. 1986. Т. 74. № 7.
- Нариньяни А.С. 1995 – Проблема понимания ЕЯ-запросов к базам данных решена // DIALOG'95. Труды Международного семинара по компьютерной лингвистике и ее приложениям. Казань, 1995.
- Окаси М. и др. 1986 – Нитиэй кикай хонгяку сисутэму Thalia-3 // Мицубиси дэнки гихо. 1986. Т. 60. № 3.
- Панов Д.Ю. 1959 – Переводная машина П.П. Троянского / Отв. ред. Д.Ю. Панов. М., 1959.
- Певзнер Б.Р. 1981 – Система машинного перевода с элементами обучения // Вопросы информационной теории и практики. № 46. М., 1991.
- Слокум Дж. 1989 – Обзор разработок по машинному переводу: история вопроса, современное состояние и перспективы развития // Новое в зарубежной лингвистике. Вып. XXIV: Компьютерная лингвистика. М., 1989.
- СМП 1962 – Сообщения по машинному переводу. Сб. 1. Таллин, 1962.
- Танака Х. – Мидзогути Ф. 1985 – Ими сери о донно сита кикай хонгяку ни цуйтэ. Юго хосики // Дзехо сери. 1985. Т. 26. № 10. (на япон. яз.).
- Утида Х. 1985 – Бун рикай ни мотодзуку кикай хонгяку сисутэму ATLAS-2 // Дзехо сери. 1985. Т. 26. № 10.
- Цейтин Г.С. 1975 – Методы синтаксического анализа, использующие предпочтение языковых конструкций (модели и эксперименты) // Международный семинар по машинному переводу. М., 1975.
- Шалапина З.М. 1975 – К проблеме построения формальной модели процесса перевода // Теория перевода и научные основы подготовки переводчиков. Материалы всесоюзной научной конференции. Ч. II. М., 1975.
- Шалапина З.М. 1980 – Автоматический перевод как моделирование переводческой деятельности человека // Международный форум по информации и документации. 1980. Т. 5. № 2.
- ALPAC 1966 – Language and machines: computers in translation and linguistics. Technical report. Washington // Automatic language processing advisory committee (ALPAC). National Academy of Sciences. Publ. № 1416.
- Amano S. 1986 – The Toshiba machine translation system // Japan computer quarterly. 1986. № 64.
- Appelo L. 1986 – A compositional approach to the translation of temporal expressions in the Rosetta system // Proceedings of COLING 86. Bonn, 1986.
- Beaven J.L. 1992 – Shake-and-bake machine translation // Proceedings of COLING 92. Nantes, 1992.
- Blanchon H. 1994 – Perspectives of DBMT for monolingual authors on the basis of LIDIA-1, an implemented mock-up // Proceedings of COLING 94. Kyoto, 1994.
- Choi K.-S. et al. 1994 – An English-to-Korean machine translator: MATES/EK // Proceedings of COLING 94. Kyoto, 1994.
- Chung Y.S. – Kunii T.L. 1986 – NARA: A two-way simultaneous interpretation system between Korean and Japanese. A methodological study // Proceedings of COLING 86. Bonn, 1986.
- Corbé M. 1960 – La machine à traduire française aurore bientôt trente ans // Automatisme. 1960. V. 5. № 3.
- Cornish T. et al. 1994 – Towards machine translation using contextual information // Proceedings of COLING 94. Kyoto, 1994.
- Gunji T. 1987 – Japanese phrase structure grammar. Dordrecht, 1987.
- Hajičová E. – Vrbová J. 1982 – On the role of the hierarchy of activation in the process of natural language understanding // Proceedings of COLING 82. Amsterdam, 1982.
- Harada T. 1986 – NEC's machine translation system "PIVOT" // Japan Computer Quarterly. 1986. № 64.
- Heylen D. et al. 1992 – Lexical functions and machine translation // Proceedings of COLING 94. Kyoto, 1994.
- Hutchins W.-J. 1986 – Machine translation: past, present, future. N.Y., 1986.
- Kanovich M.I. – Shalyapina Z.M. 1994 – The RUMORS system of Russian synthesis // Proceedings of COLING 94. Kyoto, 1994.
- Kay M. et al. 1994 – Verbmobil: a translation system for face-to-face Dialog // CSLI lecture notes. Stanford, 1994. № 33.
- Kudo I. – Nomura H. 1986 – Lexical-functional transfer: a transfer framework in a machine translation system based on LFG // Proceedings of COLING 86. Bonn, 1986.
- Kumano A. – Hirakawa H. 1994 – Building an MT dictionary from parallel texts based on linguistic and statistical information // Proceedings of COLING 94. Kyoto, 1994.
- Kushima H. – Uchida H. 1986 – Fujitsu's machine translation systems // Japan computer quarterly. 1986. № 64.
- Landsbergen J. 1987 – Montague grammar and machine translation // Linguistic theory and computer applications. L., 1987.
- LDOCE 1995 – Longman dictionary of contemporary English // Bungay (Suffolk). 1995.

- Lenke N.* 1994 – Anticipating the readers's problems and the automatic generation of paraphrases // *Proceedings of COLING 94*. Kyoto, 1994.
- Luckhart H.D. – Maas H.D.* 1983 – *SUSY – Handbuch für Transfert und Synthese*. Linguistische Arbeiten des SB-100; Neue Folge. Saarbrücken, 1983.
- Marcus M.P.* 1987 – Deterministic parsing and description theory // *Linguistic theory and computer applications*. L., 1987.
- Miura M. et al.* 1992 – Learning mechanism in machine translation System PIVOT // *Proceedings of COLING 92*. Nantes, 1992.
- Modina L.S. – Shalyapina Z.M.* 1994 – The JaRAP experimental system of Japanese-Russian automatic translation // *Proceedings of COLING 94*. Kyoto, 1994.
- Morimoto T. et al.* 1992 – A spoken language translation system: SL-TRANS2 // *Proceedings of COLING 92*. Nantes, 1992.
- Nagao M.* 1984 – A framework of a mechanical translation between Japanese and English by analogy principle // *Artificial and Human Intelligence*. Amsterdam, 1984.
- Nagao M. et al.* 1985 – The Japanese government project for machine translation // *J. assoc. comput. linguistics*. 1985. V. 11. № 2–3.
- Nagao M.* 1991 – Current state and problems of machine translation. Kyoto, 1991.
- Nirenburg S. et al.* 1986 – On knowledge-based machine translation // *Proceedings of COLING 86*. Bonn, 1986.
- Nishida F. et al.* 1986 – Construction of a modular and portable translation system // *Proceedings of COLING 86*. Bonn, 1986.
- Nomiyama H.* 1992 – Machine translation by case generalization // *Proceedings of COLING 92*. Nantes, 1992.
- Nomura H. et al.* 1986 – Translation by understanding: a machine translation system LUTE // *Proceedings of COLING 86*. Bonn, 1986.
- Nyberg E. – Mitamura T.* 1992 – The KANT System: Fast, accurate, high-quality translation in practical domains // *Proceedings of COLING 92*. Nantes, 1992.
- Papagaiti B.C.* 1986 – Word expert semantics. An interligual knowledge-based approach // *Distributed language translation*. Dordrecht: Reverton, 1986.
- Richens R.H.* 1956 – A general programme for mechanical translation between any two languages via an algebraic interlingua // *Mechanical translation*, 1956. V. 3, N 2.
- Rösner D.* 1986 – When Marico talks to Siegfried. Experience from a Japanese // German machine translation project. *Proceedings of COLING 86*. Bonn, 1986.
- Sharp R.* 1989 – CAT2 – A Formalism for multilingual machine translation // *Международный семинар по машинному переводу*. Тбилиси. 27 ноября – 2 декабря 1989 г. Тезисы докладов. М., 1989.
- Takeda K. et al.* 1992 – Shalt2 – A symmetric machine translation system with conceptual transfer // *Proceedings of COLING 92*. Nantes, 1992.
- Tomita M. – Carbonell J.G.* 1986 – Another stride towards knowledge-based machine translation // *Proceedings of COLING 86*. Bonn, 1986.
- Tsutsumi T.* 1986 – A prototype English-Japanese machine translation system for translating IBM computer manuals // *Proceedings of COLING 86*. Bonn, 1986.
- Noord G. van* 1990 – Reversible unification-based machine translation // *Proceedings of COLING 90*. Helsinki, 1990.
- Vasconcellos M. – León M.* 1985 – SPANAM and ENGSPAN: Machine translation at the Panamerican health organization // *Machine translation systems*. Cambridge, 1985.
- Vauquois B.* 1976 – Automatic translation – a survey of different approaches // *Proceedings of COLING 76*. Ottawa. V. 1. 1976.
- Vauquois B. – Boitet Ch.* 1985 – Automated translation at Grenoble university // *Machine translation systems*. Cambridge, 1985.
- Watanabe H.* 1994 – A method for distinguishing exceptional and general examples in example-based transfer systems // *Proceedings of COLING 94*. Kyoto, 1994.
- Whitelock P.* 1992 – Shake-and-bake translation // *Proceedings of COLING 92*. Nantes, 1992.
- Whitelock P. et al.* 1986 – Strategies for interactive machine translation: the experience and applications of the UMIST Japanese project // *Proceedings of COLING 86*. Bonn, 1986.