Рецензируемая книга представляет собой антологию по машинному переводу (МП), приуроченную к отмечаемому в 1999 г. полувековому юбилею этого направления. Редакторы сборника – признанные специалисты по МП, авторы оригинальных разработок. Г. Сомерс в настоящее время является редактором журнала “Machine translation”, бывший редактор этого журнала С. Ниренбург, как и Й. Уилкс, входят в Международный комитет по компьютерной лингвистике.

Сборник представляет собой переиздание 36 работ по различным аспектам МП, которые, по мнению составителей, оказали существенное влияние на техническое и идейное развитие МП, дали импульс новым направлениям исследования. Составителями отбирались только редкие, малодоступные работы, которые при этом наиболее широко цитируются. По другому критерию в сборник вошли только статьи, первоначально опубликованные на английском языке. Это объясняет, почему в нем всего одна работа российских авторов (О.С. Кулагиной и И.А. Мельчука) при большом количестве, например, японских.

Практически все авторы сборника широко известны в области МП своими пионерскими работами, и чтобы по достоинству оценить их оригинальность, читатель должен представлять технический и лингвистический контекст, в котором эти работы выходили. Предполагается также, что читатель знаком с базовыми понятиями МП, его основной проблематикой и периодизацией. В качестве учебника или исторического обзора по МП рассматривать сборник не следует. История МП, различные направления и системы достаточно подробно

освещены в другой литературе [МП 1957; АП 1971; Hutchins 1986; 2000; Слокум 1989; ИИ 1990; ONCL 2003; ACL].

Многие работы сборника, бывшие в свое время новаторскими, сейчас можно считать пройденным этапом. Однако “в построении лингвистического обеспечения систем МП бывает полезно вернуться к результатам ранее проведенных исследований в поисках ответа на сегодняшние актуальные проблемы” [Марчук 1983: 16]. В отношении же МП не перестала быть актуальной ни одна из основных проблем перевода: ни анализ текста, ни поиск переводных соответствий, ни синтез перевода. Поэтому сборник будет безусловно интересен специалистам по МП, на которых он рассчитан в первую очередь, а также всем, кто занимается смежными направлениями компьютерной лингвистики. Он также может служить прекрасным дополнением к учебной и справочной литературе по МП для студентов старших курсов и аспирантов, обучающихся по специальностям, связанным с прикладной лингвистикой и искусственным интеллектом.

Сборник покрывает период с 1949 по 1996 гг. Поскольку о исторической значимости работ последнего десятилетия говорить еще рано, в сборник они не вошли. Наиболее подробно представлено время возникновения и становления МП (до середины 1960-х) и период с начала 80-х, когда после относительного затишья 70-х началось активное развитие МП, продолжающееся до настоящего времени. В сборнике 3 раздела: исторический, теоретико-методологический и раздел, посвященный описанию конкретных систем. Как признают сами составители, такое разделение достаточно ус-

ловно, поскольку многие статьи одинаково могут быть отнесены к каждой из этих трех рубрик. Каждый раздел предварен предисловием; в конце сборника даны библиографические справки по всем работам и указатель.

Открывает сборник знаменитый “Меморандум” У. Уивера 1949 г. [МП 1957: 33–44], где впервые теоретически обоснована возможность использования ЭВМ в целях перевода. Именно с этим документом принято связывать рождение МП, хотя, как видно из статьи Э. Бута “Машинный перевод”, теоретические работы были начаты несколько раньше: первый алгоритм поиска по словарю и способ представления текста и словаря в машинном виде были предложены Бутом уже в 1947 г. Первые реально работающие системы появились на несколько лет позже и получили название МП первого поколения. Для них был характерен прямой пословно-пооборотный способ перевода, нерасчлененность алгоритмов и лингвистических знаний, ориентация на определенную пару языков и одно направление перевода (с. 21). Для преодоления этих ограничений В. Ингве в статье 1957 г. “Один подход к синтаксическому переводу” предложил трехэтапный процесс перевода, состоящий из анализа, трансфера (переводных операций) и синтеза. Описание переводного и выходного языков становятся независимыми, кроме того, разделяются лингвистическое и алгоритмическое описание. Более подробная характеристика систем МП первого и второго поколений дается в работе Б. Вокуа “Автоматический перевод – обзор различных подходов” (1976). Именно в этой статье Вокуа, по словам составителей сборника, впервые появилась широко известная теперь “пирамидальная” схема, описывающая зависимость сложности межъязыковых преобразований от глубины анализа входного текста.

Период бурного развития МП первого десятилетия, связанного с надеждами на скорые успехи, сменяется разочарованием. Й. Бар-Хиллел в критическом обзоре “Современное состояние в области автоматического перевода языков” 1960 г. убедительно показал, что полностью автоматический высококачественный перевод невозможен без учета семантики текста и знаний о мире, что в его представлении делало эту задачу практически безнадежной. Тему критики МП продолжает историческое исследование Дж. Хатчинса под заголовком “(Печально) известный доклад ALPAC” (1996), посвященное опубликованному в 1966 г. докладу Консультативного комитета по автоматизированной обработке естественных языков (ALPAC) при Национальной академии наук США. Считается, что именно заключение

этого доклада об экономической бесперспективности исследований по МП привело к сокращению финансирования работ в США, а затем и других странах. Анализируя документ в исторической перспективе, Дж. Хатчинс приходит к выводу, что скептицизм того времени о возможностях МП был вполне оправдан, хотя были проигнорированы реальные потребности промышленности и коммерции. При этом негативные последствия доклада Хатчинс считает в определенной степени преувеличенными: в 1966 г. в США не произошло полного прекращения работ по МП, несколько проектов финансировались до середины 80-х, с другой стороны Джорджтаунский, Гарвардский и некоторые другие проекты не получали государственного финансирования уже с 1963 г.

Первые работы по МП с русского, английского и других европейских языков достаточно известны. В этой связи любопытны две пионерские работы, посвященные переводу с японского языка. Сусуму Куно в статье “Предварительный подход к японско-английскому автоматическому переводу” (1961) предложил оригинальную процедуру перевода японского предложения в направлении справа налево. Такое решение продиктовано структурой японского предложения, в котором сказуемое стоит в конце. В статье “Англо-японский машинный перевод” 1959 г. японских авторов С. Такахаси, Х. Тадэнума и С. Ватанабэ представлено первое описание экспериментов по МП, проведенных в Японии на специально созданной для этих целей ЭВМ *Yamato*.

Переход к системам второго поколения сопровождался активными синтаксическими исследованиями. Два принципиально различных подхода к синтаксическому анализу представлены в работах Д. Хейса “Методы исследований в области автоматического перевода” 1963 г. [АП1971: 41–84] и И. Родес “Новый подход к машинному синтаксическому анализу русского языка” 1961 г. [АП1971: 84–101]. В первом случае, известном как метод фильтров, синтаксическая структура предложения выбирается из множества предварительно построенных вариантов анализа. Метод И. Родес – последовательный, в нем строится только одна структура, которая по необходимости исправляется методом возвратов. Другой вариант последовательного анализа представлен в работе П. Гарвина “Место звистик в алгоритме машинного перевода *Фулькрум*” 1968 г. [АП1971: 26–40].

Р. Каплан, К. Неттер, Ю. Ведыкин и А. Ценен в статье “Перевод на основе структурных соответствий” (1989) описывают схему МП, в которой нахождение переводных соответствий возложено на многоуровневый трансфер, по-

строенный на основе Лексико-функциональной грамматики Дж. Бреснан и Р. Каплана.

К 1970-м гг. все более очевидной становится ограниченность чисто синтаксического подхода и необходимость использования в переводе семантики. Новый подход оформился в модели с так называемым **семантическим языком-посредником**. Хотя на возможность такого подхода указывалось уже У. Уивером (с. 17), неисследованность семантики и сложность ее формального описания стали главным препятствием для его реализации в полном виде.

Одной из первых возможностей создания искусственного семантического языка-посредника для многоязычного МП исследовала М. Мастерман в статье "Машинный перевод на язык пиджин", 1967 г. В наиболее полном виде представление о МП как о процессе многоступенчатого преобразования входного текста в выходной через семантический уровень представлено в статье О.С. Кулагиной и И.А. Мельчука "Автоматический перевод: некоторые теоретические аспекты и общая схема системы перевода" (1967).

Среди языков-посредников выделяют языки логического и лингвистического типов. Вопрос их соотношения и возможность совместного использования рассматривает Й. Уилкс в статье "Стэнфордский проект машинного перевода" (1973). Одна из основных проблем перевода с использованием **логического языка-посредника** – проблема "непереводимого подмножества" – описана Я. Ландсбергенем в статье "Грамматика Монтегю и машинный перевод" (1987). Перевод на логический язык-посредник всех возможных значений входного и выходного языков связан с большими вычислительными трудностями. Если число переводимых значений ограничить, то может оказаться, что для некоторого подмножества выражений на языке-посреднике, получаемом при анализе, не окажется соответствий на выходном языке.

Общие проблемы, возникающие в системах МП с языком-посредником, описываются в статье К. Боте "За и против подходов с языком-посредником и трансфером в случае многоязычного машинного перевода" (1988). В их числе: собственно теоретическая возможность создания языка-посредника, трудность сохранения порядка слов и стиля в тексте перевода, рентабельность такого подхода в случае нескольких языков и направлений перевода. Из-за наличия преимуществ и слабых сторон как у подхода с языком-посредником, так и у трансферного подхода, у каждого из них были свои противники и защитники. Своего рода черту в споре между ними подводят С. Ниренбург и К. Гудман в работе "Использование зна-

чения в МП" (1990): в той или иной степени обращение к семантике присутствует почти в каждой системе, вопрос лишь в степени ее реального использования и в том, что декларируется самими создателями системы. Как показывают авторы, нередко причина неудачи семантического подхода неоправданно приписывается языку-посреднику там, где в действительности используется не язык-посредник, а трансфер (с. 288).

Важной составляющей ставшего традиционным МП с трансфером или языком-посредником является экспликация лингвистических, в том числе контрастивных знаний в виде набора правил (rule-based translation). Задачей лингвиста становится поддержание и расширение таких правил в соответствии с принципом лингвистического описания, принятым в системе. С развитием системы такие подходы сталкиваются с трудноразрешимыми сложностями, в частности, с необходимостью контроля за неограничиваемым объемом правил и невозможностью предсказания эффекта от добавления новых правил. Для преодоления противоречий традиционного подхода было предложено получать лингвистические знания не от эксперта-лингвиста, а извлекать непосредственно из текстов. Этот принцип лег в основу корпусных и статистических методов, объединяемых по тому признаку, что центральная роль в них отводится текстовым данным (data-driven approaches). К настоящему времени они заняли в МП доминирующее положение, чему немало способствовало появление больших корпусов, Интернета и дешевых компьютеров.

Первые идеи использования **корпуса текстов** высказывались С. Лэмом еще в 1961 г. В работе "Об автоматизации синтаксического анализа" он описал процедуру автоматического построения синтаксического описания на основе текста. Настоящий же расцвет корпусных методов в МП произошел сравнительно недавно, и первенство здесь отдается выдающемуся японскому исследователю М. Нагао и его работе "Машинный перевод для японского и английского языков, основанный на аналогии" 1981 г. В ней впервые предложено вместо набора переводных правил использовать параллельный корпус и тезаурус. Идея Нагао была реализована только в 90-х. По предложенной им схеме сначала в японской части корпуса ищется предложение грамматически и лексически наиболее близкое к переводимому предложению. Для оценки подобия лексических единиц используется тезаурус и специальные функции оценки "семантической близости". Затем из английской части корпуса берется перевод выбранного образца и в нем производится замена лексических единиц с со-

ответствии с тем, что есть в исходном японском предложении. Если в корпусе не обнаруживается подходящего образца, перевод оказывается невозможен. Такой подход избавляет от необходимости содержать большой набор лингвистических знаний, записанных в виде правил, а качество перевода до определенной степени можно повышать простым пополнением корпуса. Однако, как показали дальнейшие исследования, при всех своих преимуществах, без использования лингвистических правил такой подход не может достичь высокой точности перевода [Sumita, Iida 1991].

Независимо от Нагао в середине 1980-х в голландской компании BSO был предложен похожий метод, описанный в статье В. Садлера “Банк текстовых знаний: проектирование, создание и применение” (1991). Описанный “двухязычный банк знаний” сейчас назвали бы выровненным параллельным корпусом с синтаксической разметкой. В проекте BSO было предложено устанавливать соответствия между текстом и его переводом на уровне лексических единиц и групп составляющих. В отличие от М. Нагао, который говорил только о поиске переводных соответствий, у В. Садлера предполагалось извлекать из корпуса все возможные виды лингвистической информации, в том числе экстралингвистическую.

Возможность использовать лингвистические знания в их неявном виде, как они есть в парах переводных соответствий, дала начало таким направлениям как “МП, основанный на примерах” (example-based translation) и “МП с использованием банка переводных соответствий” (memory-based translation).

Корпусный подход тесно связан со **статистическим**. Указание на возможность использования статистического анализа для снятия омонимии было уже в Меморандуме Уивера (с. 16), в 1956 г. Г. Кинг в статье “Вероятностные методы машинного перевода” предложил несколько идей для разрешения текстовой неоднозначности статистическими методами, в частности указал на возможность статистического определения жанра текста, снятия частеречной неоднозначности, перевода идиом и терминологических словосочетаний (с. 38).

До 90-х методы статистического анализа в компьютерной лингвистике не имели широкого признания, что принято связывать с их негативной оценкой в работах Н. Хомского и Дж. Миллера [Chomsky, Miller 1963]. Импульс статистическому подходу к МП дала работа “Статистический подход к машинному переводу” (1990) группы авторов из ИВМ Т. J. Watson Research Center под руководством Ф. Джелинека. Она до сих пор считается хорошим введением в область статистического перевода. Группа

Джелинека изначально работала над статистическими методами распознавания речи, и именно успехи этих разработок подтолкнули к применению так называемой N-граммной модели к анализу текстов. В данной статье впервые была представлена статистическая модель именно перевода в целом, а не отдельных его этапов. Если до этого МП был жестко ориентирован на точность перевода и одновременно его формальную правильность с точки зрения выходного языка, то статистическая модель позволила гибко балансировать между этими противоречивыми требованиями, компенсируя недостаток одного другим. В описанной модели за единицу перевода принято слово, но авторы справедливо предполагают, что более адекватным будет выделение в качестве переводных единиц целых фраз. Как показывают новые статистические модели, учет дистантных зависимостей между словами или группами слов в предложении дает более точные результаты [GuoDong, KimTeng 1998; Yamada, Knight 2001].

Хотя статистические методы позволяют преодолеть некоторые проблемы традиционного подхода, они не всеисильны. На принципиальную ограниченность чистого статистического подхода при снятии некоторых видов неоднозначности указывают Ниренбург и Гудман (с. 284). Сейчас все очевиднее становится целесообразность комбинирования переводных правил со статистическими методами [OHCL 2003: 518].

Автоматический перевод устной речи – задача значительно более трудная, чем перевод текста. Представление о том, что для такого перевода достаточно добавить к системе текстового перевода модуль распознавания речи, конечно же, неверно [OHCL 2003: 521]. Как показано Д. Цудзии и М. Нагао в статье “Перевод диалога vs. перевод текста – подход, основанный на интерпретации” (1988), закономерности построения устного и письменного текста, их формальные средства слишком различны, чтобы можно было свести одну задачу к другой. Кроме того, в случае диалогической речи требуется перевод не всего текста, а только части наиболее релевантной для коммуникативных задач говорящих. Задача выявления релевантной информации может упрощаться сужением предметной области, в которой ведется диалог. Первые успешные результаты в этой области принадлежат системе перевода, созданной в японской исследовательской группе ATR. Детальное описание разработки дано в статье Т. Моримото и А. Курэмацу “Автоматический перевод речи в ATR” (1993).

Отдельная тема в МП – это **синтез текста** из внутреннего представления, полученного на

этапе перевода. В одних подходах операции анализа и синтеза текста рассматриваются как принципиально различные и несводимые друг к другу, в других синтез трактуется как обратный анализ. Аргументам за и против симметричного описания анализа и синтеза посвящена статья Й. Уилкса "Откуда я пришел: обратимость анализа и синтеза в обработке естественного языка" (1990).

Одной из сложнейших задач в МП до сих пор является снятие разного рода неоднозначности в тексте. Одним из путей если не полного решения, то упрощения этой задачи является ограничение лексики и грамматики в тексте переводимого языка. Такое ограничение достигается естественным образом в случае перевода с **подъязыка** (sublanguage). Название "подъязык" не вполне точно отражает суть явления – это не обязательно подмножество какого-либо другого языка, чаще это самостоятельный язык определенной проблемной области. Дж. Лербергер в статье "Автоматический перевод и понятие подъязыка" (1982) исследует, в какой мере один подъязык отличается от других, а также какие виды неоднозначности встречаются в подъязыке, а какие нет. Так, в приводимом в качестве примера языке авиационной документации отсутствуют вопросительные предложения, связочные предложения с обобщающим значением (*двигатель – это механическое устройство*), многие существительные (*воздух, раствор, инструмент*) не употребляются в абстрактном значении и т.д. (с. 208–210). С другой стороны, в большом количестве имеются неоднозначные последовательности типа *pump case drain fitting*. Как отмечается в [Sumita, Iida 1991: 191], наиболее просто перенастраиваться с одного подъязыка на другой позволяют корпусные методы перевода: грубо говоря, для этого достаточно лишь заменить корпус.

Другим возможным путем является искусственное **ограничение языка** (controlled language). В системе МП с ограниченным языком подаваемый на перевод текст либо изначально пишется с использованием ограниченного словаря и заданного набора грамматических конструкций, либо приводится к нужному виду редактором. Ограниченный вариант английского и метод его использования в компании XEROX описаны в статье Дж. С.Г. Эллистона "Перевод с помощью компьютера: коммерческий подход" (1979). Характерным для английского языка является запрет на конверсивное употребление слова: например, в описываемом языке *switch* можно использовать только в значении "переключатель", но не "переключать".

В МП всегда существовала проблема выбора между созданием коммерческого продукта

или инструмента для исследования свойств языка и текста. Преодолеть это противоречие пытаются Р. Джонсон и П. Уайтлок: "Система МП должна обеспечивать не только более эффективное использование человеческих возможностей, но и создавать для исследователя условия, в которых можно было бы лучше понять, какие знания необходимы для перевода" ("Машинный перевод как экспертная задача", 1985, с. 235).

До сих пор нет единства в вопросе о целесообразности полностью автоматизированного МП. Аргументы за и против МП без вмешательства человека рассматривает М. Кэй в статье "Должное место человека и компьютера при переводе естественных языков" 1980 г. Статья А. Мелби "Многоуровневые вспомогательные средства для перевода" (1982) поднимает проблему недостаточного внимания, уделяемого редакторам и переводчикам, которым приходится работать с системами МП.

Современный МП освоил перевод на уровне предложения. Однако при выходе за границы предложения или абзаца требуется учет до сих пор мало исследованных языковых правил, действующих в масштабах целого текста. Для преодоления такого рода трудностей Г. Сомерс, Д. Цудзини и Д. Джонс в работе "Машинный перевод без входного текста" 1990 г. предложили радикально новый подход к переводу. Их гипотетическая система предназначена для создания перевода, а не исходного текста, и принципиально отличается от "рабочего места переводчика", снабженного МП. Такая система позволила бы автору не писать полного текста на исходном языке, как это принято, а сразу создавать перевод с учетом правил структурирования текста, принятых в выходном языке, т.е. учитывать намерение самого автора, что в традиционном переводе возможно только в качестве гипотезы.

Создание МП как любого крупного компьютерного проекта связано с выбором средств программной реализации. Общую ситуацию описывают Р. Джонсон, М. Кинг, Л. де Томб в статье "EUROTRA: методы вычислений" (1985). Как правило в МП используется либо язык программирования общего назначения, либо создается специальный язык, ориентированный на лингвистов с разной степенью математической подготовки. Общая закономерность такова: чем более произвольна, свободна форма записи, тем более эксплицитным становится лингвистическое описание. И наоборот: большая формализованность языка ведет к лаконичности записи.

Можно заключить, что сборник отвечает задачам, поставленным составителями. Помимо возвращения в научный обиход многих ред-

ких, но влиятельных работ книга имеет и символическое значение – это своего рода дань признания и памяти исследователям, которые внесли весомый вклад в становление и развитие МП. Как видим, охвачены все важные направления в МП и основная проблематика. Из тем, отдельно никак не отраженных, можно назвать важную для МП проблему оценки качества перевода. В МП она разрабатывалась давно и имеет свою специфику по сравнению с другими направлениями компьютерной лингвистики (см. в частности о приложении к докладу ALPAC, составленном Дж. Кэрролом, оказавшем большое влияние на способ оценки многих дальнейших проектов, с. 132). К сожалению, в исторической части не представлены ранние работы по МП с китайского языка (Э. Райфлер, Г. Кинг), хотя успехи в этой области были сопоставимы с успехами русско-английского перевода [Hutchins 1986, гл. 4]. Возникает также естественное желание увидеть в сборнике больше работ отечественных авторов. По словам составителей, предполагается продолжение данного сборника, освещающее работы национальных школ МП, и можно надеяться, что российские разработки будут представлены полнее. Однако очевидно, что необходима и отечественная антология подобного рода, отражающая основные этапы и достижения советских и российских исследований в области МП.

СПИСОК ЛИТЕРАТУРЫ

- АП 1971 – Автоматический перевод. Сб. статей / Ред. О.С. Кулагина и И.А. Мельчук. М., 1971.
- ИИ 1990 – Искусственный интеллект. 1: Системы общения и экспертные системы / Ред. Э.В. Попов. М., 1990.
- Марчук 1983 – Ю.Н. Марчук. Проблемы машинного перевода. М., 1983.
- МП 1957 – Машинный перевод. Сб. статей / Под ред. П.С. Кузнецова. М., 1957.

- Слокум 1989 – Дж. Слокум. Обзор разработок по машинному переводу: история вопроса, современное состояние и перспективы развития // Новое в зарубежной лингвистике. Вып. XXIV: Компьютерная лингвистика. М., 1989.
- ACL – <http://www.aclweb.org/anthology> (Труды международных конференций COLING и ACL, семинаров по МП, а также тексты журнала Computational Linguistics).
- Chomsky, Miller 1963 – N. Chomsky, G. Miller. Introduction to the formal analysis of natural languages // Handbook of Mathematical psychology 2 / Ed. by R.D. Luce, R. Bush, E. Galanter. New York, 1963.
- GuoDong, KimTeng 1998 – GuoDong Zhou, KimTeng Lua. Word association and MI-Trigger-based language modeling // Proceedings of the 17th international conference on computational linguistics. Montréal, 1998.
- Hutchins 1986 – J. Hutchins. Machine translation: past, present, future. Chichester, 1986. См. также: <http://ourworld.compuserve.com/homepages/WJHutchins/>.
- Hutchins 2000 – Early years in machine translation. Memoirs and biographies of pioneers / Ed. by W.J. Hutchins // Studies in the history of the language sciences. Amsterdam; Philadelphia, 2000.
- ONCL 2003 – The Oxford handbook of computational linguistics / Ed. by R. Mitkov. Oxford, 2003.
- Sumita, Iida 1991 – Eiichiro Sumita, Hitoshi Iida. Experiments and prospects of example-based machine translation // Proceedings of the 29th annual meeting of the Association for computational linguistics (ACL-91). Berkeley, USA, 1991.
- Yamada, Knight 2001 – K. Yamada, K. Knight. A syntax-based statistical translation model // Proceedings of the conference of the association for computational linguistics (ACL). Toulouse, 2001.