Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

**Анна Юрьевна Филиппович**

# СИСТЕМЫ МАШИННОГО ПЕРЕВОДА

## Лекции по дисциплине "Лингвистическое обеспечение АСОИУ"

*Модуль 2*
*Лингвистическое программное обеспечение*

Москва, 2012

# Системы машинного перевода

### Определение

***Машинный перевод (МП)*** – автоматический перевод, перевод текстов с одного языка на другой с помощью автоматических устройств.

***Машинный перевод*** – выполняемое на компьютере действие по преобразованию текста на одном естественном языке в эквивалентный по содержанию текст на другом языке, а также результат такого действия.

## История развития систем машинного перевода

### 40-е: первые шаги

История машинного перевода как научно-прикладного направления началась в конце 40-х годов прошлого века (если не считать механизированное переводное устройство П. П. Смирнова-Троянского, своего рода лингвистический арифмометр, изобретенный в 1933 году). Теоретической основой начального (конец 1940-х – начало 1950-х годов) периода работ по машинному переводу был взгляд на язык как кодовую систему. Пионерами МП были математики и инженеры. Описания их первых опытов, связанных с использованием только что появившихся ЭВМ для решения криптографических задач, были опубликованы в США в конце 1940-х годов. Датой рождения машинного перевода как исследовательской области обычно считают март 1947; именно тогда специалист по криптографии Уоррен Уивер в своем письме Норберту Винеру впервые поставил задачу машинного перевода, сравнив ее с задачей дешифровки.

Тот же Уивер после ряда дискуссий составил в 1949 г. меморандум, в котором теоретически обосновал принципиальную возможность создания систем машинного перевода. У. Уивер писал: «I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text» («У меня перед глазами текст, написанный по-русски, но я собираюсь сделать вид, что на самом деле он написан по-английски и закодирован при помощи довольно странных знаков. Все, что мне нужно, — это взломать код, чтобы извлечь информацию, заключенную в тексте»). Аналогия между переводом и дешифрованием была естественной в контексте послевоенной эпохи, если учитывать успехи, которых достигла криптография в годы Второй мировой войны.

Идеи Уивера легли в основу подхода к МП, основанного на концепции interlingva: стадия передачи информации разделена на два этапа. На первом этапе исходное предложение переводится на язык-посредник (созданный на базе упрощенного английского языка), а затем результат этого перевода представляется средствами выходного языка.

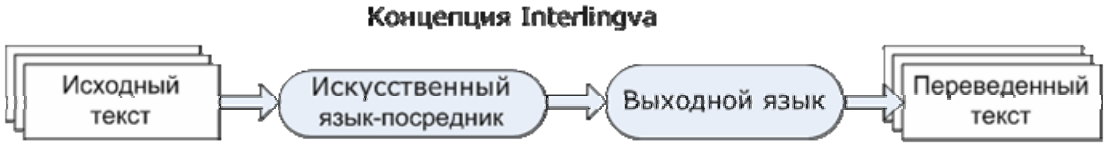


Рис. 1. Концепция «Интерлингва».

Меморандум Уивера вызвал самый живой интерес к проблеме МП. В 1948 г. А. Бут и Ричард Риченс (Richard Richens) произвели некоторые предварительные эксперименты (так, Риченс разработал правила разбиения словоформ на основы и окончания). Вскоре началось финансирование исследований. На ранних этапах разработка МП активно

поддерживалась военными, при этом в США основное внимание уделялось русско-английскому направлению, а в СССР — англо-русскому.

Помимо очевидных практических нужд важную роль в становлении машинного перевода сыграло то обстоятельство, что предложенный в 1950 г. английским математиком А. Тьюрингом знаменитый тест на разумность («тест Тьюринга») фактически заменил вопрос о том, может ли машина мыслить, на вопрос о том, может ли машина общаться с человеком на естественном языке таким образом, что тот не в состоянии будет отличить ее от собеседника-человека. Тем самым вопросы компьютерной обработки естественно-языковых сообщений на десятилетия оказались в центре исследований по кибернетике (а впоследствии по искусственному интеллекту), а между математиками, программистами и инженерами-компьютерщиками с одной стороны и лингвистами – с другой установилось продуктивное сотрудничество.

В 1952 г. состоялась первая конференция по МП в Массачусетском технологическом университете, а в 1954 г. в Нью-Йорке была представлена первая система МП — IBM Mark II, разработанная компанией IBM совместно с Джорджтаунским университетом (это событие вошло в историю как Джорджтаунский эксперимент). Была представлена очень ограниченная в своих возможностях программа (она имела словарь в 250 единиц и 6 грамматических правил), осуществлявшая перевод с русского языка на английский. В том же 1954-м первый эксперимент по машинному переводу был осуществлен в СССР И. К. Бельской (лингвистическая часть) и Д. Ю. Пановым (программная часть) в Институте точной механики и вычислительной техники Академии наук СССР, а первый промышленно пригодный алгоритм машинного перевода и система машинного перевода с английского языка на русский на универсальной вычислительной машине были разработаны коллективом под руководством Ю. А. Моторина. После этого работы начались во многих информационных институтах, научных и учебных организациях страны. Казалось, что создание систем качественного автоматического перевода вполне достижимо в пределах нескольких лет (при этом акцент делался на развитии полностью автоматических систем, обеспечивающих высококачественные переводы; участие человека на этапе постредактирования расценивалось как временный компромисс). Профессиональные переводчики всерьез опасались в скором времени остаться без работы...

## 50-е: первое разочарование

К началу 50-х годов целый ряд исследовательских групп в США и в Европе работали в области МП. В эти исследования были вложены значительные средства, однако результаты очень скоро разочаровали инвесторов. Одной из главных причин невысокого качества МП в те годы были ограниченные возможности аппаратных средств: малый объем памяти при медленном доступе к содержащейся в ней информации, невозможность полноценного использования языков программирования высокого уровня. Другой причиной было отсутствие теоретической базы, необходимой для решения лингвистических проблем, в результате чего первые системы МП сводились к пословному (word-to-word) переводу текстов без какой-либо синтаксической (а тем более смысловой) целостности.

В 1959 г. философ Й. Бар-Хиллел (Yohoshua Bar-Hillel) выступил с утверждением, что высококачественный полностью автоматический МП (FAHQMT) не может быть достигнут в принципе. В качестве примера он привел проблему нахождения правильного перевода для слова pen в следующем контексте: John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy (Джон искал свою игрушечную коробку. Наконец он ее нашел. Коробка была в манеже. Джон был очень счастлив). Pen в данном случае должно переводиться не как «ручка» (инструмент для письма), а как «детский манеж» (play-pen). Выбор того или иного перевода в этом случае и в ряде других

обусловлен знанием внеязыковой действительности, а это знание слишком обширно и разнообразно, чтобы вводить его в компьютер. Однако Бар-Хиллел не отрицал идею МП как таковую, считая перспективным направлением разработку машинных систем, ориентированных на использование их человеком-переводчиком (своего рода «человеко-машинный симбиоз»).

Это выступление самым неблагоприятным образом отразилось на развитии МП в США. В 1966 г. специально созданная Национальной Академией наук комиссия ALPAC (Automatic Language Processing Advisory Committee), основываясь в том числе и на выводах Бар-Хиллела, пришла к заключению, что машинный перевод нерентабелен: соотношение стоимости и качества МП было явно не в пользу последнего, а для нужд перевода технических и научных текстов было достаточно человеческих ресурсов. За докладом ALPAC последовало сокращение финансирования исследований в области МП со стороны правительства США — и это несмотря на то, что в то время как минимум три различные системы МП регулярно использовались рядом военных и научных организаций (в числе которых ВВС США, Комиссия США по ядерной энергии, Центр Евроатома в Италии).

### 60-е: низкий старт

Следующие десять лет разработка систем МП осуществлялась в США университетом Brigham Young University в Прово, штат Юта (ранние коммерческие системы WEIDNER и ALPS) и финансировалась Мормонской церковью, заинтересованной в переводе Библии; в Канаде группами исследователей, в числе которых TAUM в Монреале с ее системой METEO; в Европе — группами GENA (Гренобль) и SUSY (Саарбрюкен). Особого упоминания заслуживает работа в этой области отечественных лингвистов, таких, как И. А. Мельчук и Ю. Д. Апресян (Москва), результатом которой стал лингвистический процессор ЭТАП. В 1960 г. в составе Научно-исследовательского института математики и механики в Ленинграде была организована экспериментальная лаборатория машинного перевода, преобразованная затем в лабораторию математической лингвистики Ленинградского государственного университета.

### 70-80-е: новый импульс

Новый подъем исследований в области МП начался в 1970-х годах и был связан с серьезными достижениями в области компьютерного моделирования интеллектуальной деятельности. Соответствующая область исследований, возникшая несколько позже МП (датой ее рождения обычно считают 1956 г.), получила название искусственного интеллекта, а создание систем машинного перевода было осмыслено в 1970-е годы как одна из частных задач этого нового исследовательского направления.

При этом несколько сместились акценты: исследователи теперь ставили целью развитие «реалистических» систем МП, предполагавших участие человека на различных стадиях процесса перевода. Системы МП из «врага» и «конкурента» профессионального переводчика превращаются в незаменимого помощника, способствующего экономии времени и человеческих ресурсов.

За период 1978-93 гг. в США на исследования в области МП истрачено 20 миллионов долларов, в Европе — 70 миллионов, в Японии — 200 миллионов.

Можно выделить два основных стимула к развитию работ по машинному переводу в современном мире. Первый – собственно научный; он определяется комплексностью и сложностью компьютерного моделирования перевода. Как вид языковой деятельности перевод затрагивает все уровни языка – от распознавания графем (и фонем при переводе устной речи) до передачи смысла высказывания и текста. Кроме того, для перевода характерна обратная связь и возможность сразу проверить теоретическую гипотезу об

устройстве тех или иных языковых уровней и эффективности предлагаемых алгоритмов. Эта характеристическая черта перевода вообще и машинного перевода в частности привлекает внимание теоретиков, в результате чего продолжают возникать все новые теории автоматизации перевода и формализации языковых данных и процессов. Вместе с тем разработки в области МП стимулировали развитие не только лингвистики. Результаты работ по МП способствовали началу и развитию исследований и разработок в области автоматизации информационного поиска, логического анализа естественно-языковых текстов, экспертных систем, способов представления знаний в вычислительных системах и т.д.

Второй стимул – социальный, и обусловлен он возрастающей ролью самой практики перевода в современном мире как необходимого условия обеспечения межъязыковой коммуникации, объем которой возрастает с каждым годом. Другие способы преодоления языковых барьеров на пути коммуникации – разработка или принятие единого языка, а также изучение иностранных языков – не могут сравниться с переводом по эффективности. С этой точки зрения можно утверждать, что альтернативы переводу нет, так что разработка качественных и высокопроизводительных систем машинного перевода способствует разрешению важнейших социально-коммуникативных задач.

Одной из новых разработок этого периода стала технология TM (translation memory), работающая по принципу накопления: в процессе перевода сохраняется исходный сегмент (предложение) и его перевод, в результате чего образуется лингвистическая база данных; если идентичный или подобный исходному сегмент обнаруживается во вновь переводимом тексте, он отображается вместе с переводом и указанием совпадения в процентах. Затем переводчик принимает решение (редактировать, отклонить или принять перевод), результат которого сохраняется системой. А в конечном итоге «не нужно дважды переводить одно и то же предложение!». В настоящее время разработчиком известной коммерческой системы, основанной на технологии TM, является система TRADOS (основана в 1984 г.).

В СССР с середины 70-х годов были созданы промышленные системы машинного перевода с английского языка на русский АМПАР (на основе исследований и разработок коллектива Ю. А. Моторина), с немецкого языка на русский НЕРПА, с французского языка на русский ФРАП, автоматические терминологические словари в помощь человеку-переводчику. Система АМПАР длительное время находилась в промышленной эксплуатации; впоследствии на ее базе были созданы более эффективные системы МП для персональных компьютеров семейства СПРИНТ; была также разработана система МП с русского языка на английский АСПЕРА. На этих разработках основываются такие системы машинного перевода, как Stylus, Socrat и другие.

### От 90-х к XXI веку

90-е годы принесли с собой бурное развитие рынка ПК (от настольных до карманных) и информационных технологий, широкое использование сети Интернет (которая становится все более интернациональной и многоязыкой). Все это сделало возможным, а главное востребованным, дальнейшее развитие систем МП. Появляются новые технологии, основанные на использовании нейронных сетей, концепции коннекционизма, статистических методах.

В настоящее время несколько десятков компаний занимаются разработкой коммерческих систем МП, в их числе: Systran, IBM, L&H (Lernout & Hauspie), Language Engineering Corporation, Transparent Language, Nova Incorporated, Trident Software, Atril, TRADOS, Caterpillar Co., LingoWare; Ata Software; Lingvistica b.v. и др. В настоящее время в Российской Федерации продолжаются в незначительных масштабах некоторые работы по системам МП, основанным на подходе «текст-смысл-текст», не всегда явно проговариваемым лозунгом которого в момент обоснования этого подхода в 1960-х годов

был «машинный перевод без перевода, без машин, без алгоритмов». Идея подхода заключалась в том, что от лингвиста требуется только декларативное описание фактов языка (т.е. лингвистическая теория, претендующая, правда, на особую точность и формализованность), а алгоритмы перевода составят программист и математик. В рамках этих исследований были получены значительные теоретико-лингвистические результаты (в частности, создана теория так называемых лексических функций, нашедшая применение в лексикографии), однако для создания практических систем подобного рода подход оказался недостаточно эффективным. Все практические системы без исключения используют идею переводных соответствий, т.е. в их основе лежит модель «текст-текст», и они реализуют краткую схему перевода. Неизмеримо выросшие за последние десятилетия возможности вычислительной техники и новые программистские подходы никак не могут помочь реализовать идеи анализа и синтеза, основанные на приоритете выявления только синтаксической структуры с последующим переходом к смыслу.

За рубежом эксплуатируется целый ряд систем машинного перевода. Наиболее известной из их числа является система Systran, разработанная и поддерживаемая компанией Systran Software Inc, используемая службой машинного перевода при комиссии Европейского союза.

Появилась возможность воспользоваться услугами автоматических переводчиков непосредственно в Сети: www.alphaworks.ibm.com/aw.nsf/html/mt; www.freetranslation.com; www.transtlate.ru; www.logomedia.net/text.asp; www.foreignword.com/Tools/transnow.htm; babelfish.altavista.com/translate.dyn; infinit.reverso.net/traduire.asp; www.t-mail.com.

С начала 1990-х годов на рынок систем ПК выходят отечественные разработчики.

В июле 1990 года на выставке PC Forum в Москве была представлена первая в России коммерческая система машинного перевода под названием PROMT (PROgrammer's Machine Translation). В 1991 г. было создано ЗАО «ПРОект МТ», и уже в 1992 г. компания «ПРОМТ» выиграла конкурс NASA на поставку систем МП (ПРОМТ была единственной неамериканской фирмой на этом конкурсе).

Несмотря на такую долгую историю, фактически всеми системами осуществляется перевод только на уровне поверхностного синтаксиса, поскольку еще не разработаны (по всей видимости) эффективные модели формального представления смысла, носителем которого должен выступать язык-посредник – интерлингва, хотя для отдельных узких отраслей такие модели строятся (например, METEO и LingoWare). Специалисты связывают построение адекватных систем МП с развитием искусственного интеллекта: машина сможет переводить с одного языка на другой, когда научится думать, как человек.

Другой путь совершенствования МП, более доступный на современном этапе, – составить корпус соответствий на двух языках. Можно предположить, что такие работы ведутся, и многими разными командами, но их действия не скоординированы, и потому результат слишком мал.

Критики современных систем МП полагают, что установка на жанровую ограниченность (научить машину сначала понимать совсем простые, специально отобранные тексты) на практике привела к тому, что задача моделирования естественного языка фактически уступила место задаче моделирования ограниченных (и крайне примитивных) подъязыков отдельных отраслей знания. При этом наилучшего результата на этом пути, как известно, достигла канадская система TAUM-METEO, отлично выполняющая задачу англо-французского перевода сводок погоды. 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.

Существующий в настоящее время «словоцентрический» подход (когда машина выбирает и переводит главным образом отдельные слова) объясняется тем, что выделяется то, что легко выделить (слова разделены пробелами), и, соответственно, это

переводится. Однако человек (в том числе тот, который занимается переводом) имеет дело с текстом, когда отдельное предложение приобретает смысл как часть более широкого контекста: соседние предложения определяют и объясняют многие невыраженные или неоднозначные элементы каждого отдельного высказывания. На настоящем же этапе часто самыми удобными для понимания оказываются такие системы МП, которые выполняют перевод пословно: фраза корявая, но видно, как она получилась, и, если есть поддержка в виде знания исходного языка, легко догадаться, что же было в оригинале, и увидеть, какие слова переведены неверно. Те системы, которые переводят текст пословно, зачастую оказываются удобнее: видно, откуда фраза взялась. Если хотя бы поверхностно знать язык оригинала, можно понять, что же было в первоначальном варианте, и какие слова переведены неверно. Системы МП, которые обрабатывают фразу синтаксически, избегая «корявости», часто выдают гладкие, но совершенно невразумительные переводы.

## Технология Translation memory (TM)

TM – это база данных, где хранятся выполненные переводы. Технология TM работает по принципу накопления: в процессе перевода в TM сохраняется исходный сегмент (предложение) и его перевод. При обработке нового текста, поступившего на перевод, система сравнивает каждое его предложение с сохраненными в базе сегментами. Если идентичный или подобный исходному сегмент найден, то перевод этого сегмента отображается вместе с переводом и указанием совпадения в процентах. Слова и фразы, которые отличаются от сохраненного текста, выделяются подсветкой. Таким образом, переводчику остается перевести только новые сегменты и отредактировать частично совпадающие. Каждое изменение или новый перевод сохраняются в TM. А в результате нет необходимости дважды переводить одно и то же предложение.


Рис. 2. Технология TM.

С другой стороны, при работе с крупными проектами переводчик сталкивается с проблемой согласованного применения терминологического глоссария в ходе длительного проекта или быстрого повторного использования ранее переведенного текста. По своей природе подобные рутинные задачи сравнительно легко (в отличие от машинного перевода) формализуются и программируются.

Каждая запись базы данных TM представляет собой единицу (предложение или абзац) параллельных текстов (как правило, на двух языках). Такая база данных хранит предыдущие переводы с целью их возможного повторного использования и решения задач быстрого поиска по содержимому. Несмотря на то что программы, оснащенные памятью перевода, называются системами автоматизированного перевода (CAT, computer-aided/assisted translation), их не следует путать с программами машинного перевода (machine translation) – память перевода ничего не переводит сама по себе, в то время как машинный перевод основан на генерации переводов по результатам грамматического разбора исходного текста.

Как правило, запись памяти перевода состоит из двух сегментов: на исходном (source) и конечном (target) языках. Если идентичный (или похожий) сегмент на исходном

языке встречается в тексте, сегмент на конечном языке будет найден в памяти перевода и предложен переводчику в качестве основы для нового перевода. Автоматически найденный текст может быть задействован как есть, отредактирован или полностью отвергнут. Большинство программ используют алгоритм нечеткого соответствия (fuzzy matching), существенно улучшающий их функциональные возможности, поскольку в этом случае можно находить предложения, лишь отдаленно напоминающие искомые фразы, но, тем не менее, пригодные для последующего редактирования.

Преимущества от использования такого программного обеспечения поначалу могут быть неочевидны – однако по мере наполнения базы данных результаты автоматической подстановки основ для перевода будут становиться все более точными и регулярными.

Архитектура автоматизированной системы и ее функциональные возможности могут различаться. Средства поиска могут работать как с целыми сегментами, так и с отдельными словами или фразами, позволяя переводчику выполнять терминологический поиск. В систему также включают отдельную программу для работы с глоссарием, содержащим утвержденные для применения в проекте термины. Некоторые системы работают с программами машинного перевода. Основной рабочий интерфейс либо встраивается непосредственно в имеющийся текстовый процессор, такой как Word, либо представляет собой отдельный редактор. В состав системы обязательно включают фильтры для импорта-экспорта файлов различных форматов. Кроме того, многие системы, если не все, имеют средство для добавления в память перевода сегментов из, как правило, имеющихся у переводчика старых переведенных файлов.

То, что применимо к понятию «обучение языку», применимо и к «Translation Memories».

- «Пустая» система запоминает термины и предложения.
- Строится «память переводов» – Translation Memory (TM).
- TM становится «языковой памятью» по продукту или по деятельности компании в целом.

Системы TM: SDLX, TRADOS, Deja Vu, Star Transit, Trans Suite 2000, WordFast, WordFisher, ACROSS.

## Комбинированные системы

Технологии МП и TM друг друга дополняют, но никак не дублируют. Система МП готова к использованию сразу после установки (хотя это не исключает того, что в процессе работы пользователю захочется что-то изменить в словаре, алгоритмах перевода и т.п.). Систему TM необходимо специально настраивать на перевод текстов в какой-то конкретной области, и чем больше эти тексты друг на друга похожи (например, такая система используется для перевода стандартных договоров), тем меньше времени требуется для настройки.

В связи с этим совершенно логично появление гибридов – example-based machine translation – программ, объединяющих системы машинного перевода и TM (например, компания «ПРОМТ» создала интегрированную технологию PROMT Term и PROMT For TRADOS, которая объединяет систему TM TRADOS и систему машинного перевода – PROMT XT Professional). PROMT For TRADOS (P4T) предназначена для интеграции системы машинного перевода PROMT и системы TM TRADOS:

- перевод в системе TRADOS;
- перевод в системе PROMT не найденных в TM сегментов;
- вставка переведенных PROMT сегментов в TM.

Схема автоматизированной цепочки перевода на основе интегрированной технологии PROMT-TRADOS

Применение интегрированной технологии делает процесс перевода больших массивов документации управляемым и повышает его экономическую эффективность.

Пример реализации проектов с применением интегрированной технологии PROMT-TRADOS.

Предположим, необходимо осуществить перевод инструкции к мини-АТС.

1. На первом этапе применяется программа PROMT TerM. Документы анализируются, и выявляется основная терминология, которая заносится в словари системы машинного перевода PROMT.

2. Выполняется машинный перевод (МП) с подключенным словарем, продолжается терминологическая работа по коррекции словаря.

3. Результаты МП корректируются и заносятся в ТМ переводимого документа.

4. Таким образом пользователь получает:

- терминологический словарь;
- переведенный документ;
- соответствующую переведенному документу ТМ, которая может быть использована в дальнейшей работе с документами подобного рода.

## Системы МП: этапы анализа текста

В ее основе работы системы МП лежит алгоритм перевода – последовательность однозначно и строго определенных действий над текстом для нахождения соответствий в данной паре языков L1 – L2 при заданном направлении перевода (с одного конкретного языка на другой). Обычные словари и грамматики разных языков не применимы для машинного перевода, так как описывают значения слов и грамматические закономерности в нестрогой форме, никак не приемлемой для «машинного» использования. Следовательно, нужна формальная грамматика языка, т.е. логически непротиворечивая и явно выраженная (безо всяких подразумеваний и недомолвок). Как только начали появляться формальные описания различных областей языка – прежде всего морфологии и синтаксиса, – наметился прогресс и в разработке систем автоматического перевода. Чтобы успешно работать, система машинного перевода включает в себя, во-первых, двуязычные словари, снабженные необходимой информацией (морфологической, относящейся к формам слова, синтаксической, описывающей способы сочетания слов в предложении, и семантической, т.е. отвечающей за смысл), а во-вторых – средства грамматического анализа, в основе которых лежит какая-нибудь из формальных, т.е. строгих, грамматик. Наиболее распространенной является следующая последовательность формальных операций, обеспечивающих анализ и синтез в системе машинного перевода.

1. На первом этапе осуществляется ввод текста и поиск входных словоформ (слов в конкретной грамматической форме, например дательного падежа множественного числа) во входном словаре (словаре языка, с которого производится перевод) с сопутствующим морфологическим анализом, в ходе которого устанавливается принадлежность данной словоформы к определенной лексеме (слову как единице словаря). В процессе анализа из формы слова могут быть получены также сведения, относящиеся к другим уровням организации языковой системы, например, каким членом предложения может быть данное слово. В школьном грамматическом разборе предложения мы опираемся и на значения слов, составляющих предложение (например, отыскивая подлежащее, задаем вопрос: о чем говорится в предложении?). Для машины же совмещение двух этих операций – и грамматического разбора, и обращения к смыслу слов – задача трудная. Лучше сделать синтаксический анализ не зависящим от смысла слов, а словарь использовать на других этапах перевода.

Что такое независимый синтаксический анализ, можно понять, если попытаться разобрать фразу, из которой «убраны» значения конкретных слов. Блестящим образцом фразы такого рода является придуманное академиком Л. В. Щербой предложение: Глокая куздра штетко будланула бокра и кудрячит бокрёнка. Бессмысленная фраза? Как будто да: в русском языке нет слов, из которых она состоит (кроме союза и). И все же в какой-то

степени мы ее понимаем: «куздра» – это существительное (мы даже можем предположить, что оно обозначает какое-то животное), «глокая» – определение к нему, «будланула» – глагол-сказуемое (похожий на толканула, боднула), «штетко» – скорее всего, обстоятельство образа действия (что-то вроде сильно, резко), «бокра» – это прямое дополнение («будланула» кого? – «бокра») и т. д.

То есть машина осуществляет синтаксический анализ предложения без опоры на значения составляющих его слов, с использованием информации только об их грамматических свойствах. В результате синтаксического анализа возникает синтаксическая структура, которая изображается в виде дерева зависимостей: «корень» – сказуемое, а «ветви» – синтаксические отношения его с зависимыми словами. Каждое слово предложения записывается в своей словарной форме, а при ней указываются те грамматические характеристики, которыми обладает это слово в анализируемом предложении.

2. Следующий этап включает в себя перевод идиоматических словосочетаний, фразеологических единств или штампов данной предметной области (например, при англо-русском переводе обороты типа in case of, in accordance with получают единый цифровой эквивалент и исключаются из дальнейшего грамматического анализа); определение основных грамматических (морфологических, синтаксических, семантических и лексических) характеристик элементов входного текста (например, числа существительных, времени глагола, их роли в данном предложении и пр.), производимое в рамках входного языка; разрешение неоднозначности (скажем, англ. round может быть существительным, прилагательным, наречием, глаголом или же предлогом); анализ и перевод слов. Обычно на этом этапе однозначные слова отделяются от многозначных (имеющих более одного переводного эквивалента в выходном языке), после чего однозначные слова переводятся по спискам эквивалентов, а для перевода многозначных слов используются так называемые контекстологические словари, словарные статьи которых представляют собой алгоритмы запроса к контексту на наличие/отсутствие контекстных определителей значения.

3. Окончательный грамматический анализ, в ходе которого доопределяется необходимая грамматическая информация с учетом данных выходного языка (например, при русских существительных типа сани, ножницы глагол должен стоять в форме множественного числа, притом, что в оригинале может быть и единственное число).

4. Синтез выходных словоформ и предложения в целом на выходном языке. Здесь не получится обойтись простым переводом «узлов» дерева на другой язык. Синтаксис каждого языка устроен на свой лад: то, что в русском предложении – подлежащее, в другом языке может (или должно) быть выражено дополнением, а дополнение, наоборот, должно преобразоваться в подлежащее; то, что в одном языке обозначается группой слов, переводится на другой всего одним словом и т. д. Так, при переводе русской фразы «У меня была интересная книга» на английский язык глагол «быть» надо перевести глаголом to have – «иметь», сочетание «у меня» преобразовать в подлежащее I («я»), а слово «книга», которое в русском языке – подлежащее, по-английски должно стать прямым дополнением: I had an interesting book (буквально: «Я имел интересную книгу»). В связи с этим в машинную память помимо наборов синтаксических правил для каждого языка «вкладывают» и правила преобразования синтаксических структур. К этому добавляют правила перехода от уже преобразованной структуры к предложению того языка, на который делается перевод. Такой переход от структуры к реальному предложению называется синтаксическим синтезом.

В зависимости от особенностей морфологии, синтаксиса и семантики конкретной языковой пары, а также направления перевода общий алгоритм перевода может включать и другие этапы, а также модификации названных этапов или порядка их следования, но вариации такого рода в современных системах, как правило, незначительны. Анализ и синтез могут производиться как пофразно, так и для всего текста, введенного в память

компьютера; в последнем случае алгоритм перевода предусматривает определение так называемых анафорических связей (такова, например, связь местоимения с замещаемым им существительным – скажем, местоимения им со словом местоимения в самом этом пояснении в скобках).

Для решения проблемы многозначности слова используется анализ контекста. Дело в том, что каждое из нескольких значений многозначного слова в большинстве случаев реализуются в своем наборе контекстов. То есть у каждого из «конкурирующих» (при интерпретации) значений – свой набор контекстов. И именно вот эта зависимость значения от окружения позволяет слушающему понять высказывание правильно. Для правильного понимания высказывания необходимо в полной мере учитывать также правила обусловленности выбранного значения лексическим окружением (действующие при «фразеологической» интерпретации слова), правила обусловленности выбранного значения семантическим контекстом (так называемые законы семантического согласования) и правила обусловленности выбранного значения грамматическим (морфолого-синтаксическим) контекстом. То есть для решения проблемы «моносемизации» слов при автоматическом переводе основой служит изучение и тщательное описание закономерностей лексической, семантической и грамматической сочетаемости. При этом правила такой сочетаемости достаточно подробно описываются в словарях – а именно, (а) с мощным охватом лексики, но весьма бегло и нетщательно, а также весьма имплицитно это делается в традиционной лексикографии; и, с другой стороны, (б) в выборочном порядке (со слабым охватом лексики), но зато весьма аккуратно и тщательно, и довольно-таки эксплицитно это делается в работах по «толково-комбинаторной» лексикографии (последних сорока лет).

## Недостатки систем машинного перевода

Действующие системы машинного перевода, как правило, ориентированы на конкретные пары языков (например, французский и русский или японский и английский) и используют, как правило, переводные соответствия либо на поверхностном уровне, либо на некотором промежуточном уровне между входным и выходным языком. Качество машинного перевода зависит от объема словаря, объема информации, приписываемой лексическим единицам, от тщательности составления и проверки работы алгоритмов анализа и синтеза, от эффективности программного обеспечения. Современные аппаратные и программные средства допускают использование словарей большого объема, содержащих подробную грамматическую информацию. Информация может быть представлена как в декларативной (описательной), так и в процедурной (учитывающей потребности алгоритма) форме.

В практике переводческой деятельности и в информационной технологии различаются два основных подхода к машинному переводу. С одной стороны, результаты машинного перевода могут быть использованы для поверхностного ознакомления с содержанием документа на незнакомом языке. В этом случае он может использоваться как сигнальная информация и не требует тщательного редактирования. Другой подход предполагает использование машинного перевода вместо обычного «человеческого». Это предполагает тщательное редактирование и настройку системы перевода на определенную предметную область. Здесь играют роль полнота словаря, ориентированность его на содержание и набор языковых средств переводимых текстов, эффективность способов разрешения лексической многозначности, результативность работы алгоритмов извлечения грамматической информации, нахождения переводных соответствий и алгоритмов синтеза. На практике перевод такого типа становится экономически выгодным, если объем переводимых текстов достаточно велик (не менее нескольких десятков тысяч страниц в год), если тексты достаточно однородны, словари

системы полны и допускают дальнейшее расширение, а программное обеспечение удобно для постредактирования.