**УДК 81.282.3**

# НЕКОТОРЫЕ АСПЕКТЫ СТАНОВЛЕНИЯ И РАЗВИТИЯ МАШИННОГО ПЕРЕВОДА

**И.А. Погуляев[1], Е.П. Игнатьева[2]**
Иркутский национальный исследовательский технический университет,
664074, Россия, г. Иркутск, ул. Лермонтова, 83.
В статье рассмотрены этапы развития машинного перевода: от первых экспериментов до сегодняшних дней; рассказывается о видах машинного перевода и об учёных, посвятивших жизнь улучшению коммуникации между людьми.
*Ключевые слова: машинный перевод*; *естественный язык; компьютерная программа; переводчик; язык.*

**SOME ASPECTS OF MACHINE TRANSLATION EVOLUTION**
**I. Pogulyaev, E. Ignatyeva**
Irkutsk National Research Technical University,
83 Lermontov Str., Irkutsk, Russia, 664074
The article describes stages of development of machine translation: from the first experiments until today. Besides, the article focuses on types of machine translation and scientists who devoted their lives to improving communication between people.
*Keywords: machine translation, natural language, computer program, translator, language.*

Цель настоящей статьи состоит в том, чтобы показать эволюцию машинного перевода, который в современных условиях развития лингвистической науки играет одну из важных ролей.

Машинный перевод – процесс перевода текстов (письменных, а в идеале и устных) с одного естественного языка на другой с помощью специальной компьютерной программы [1]. Также называется направление научных исследований, связанных с построением подобных систем. Идея о том, чтобы облегчить процесс перевода с одного естественного языка на другой, давно заботит учёных, но, так как естественный язык как высокоорганизованная система плохо поддаётся формализации, работа в этом направлении ведётся очень давно.

Таким образом, история машинного перевода насчитывает уже 200 лет и своими корнями уходит в XIX в. Впервые мысль о возможности машинного перевода высказал Чарльз Бэббидж (1791-1871), разработавший в 1836-1848 гг. проект цифровой аналитической машины – механического прототипа электронных цифровых вычислительных машин, появившихся через 100 лет. Идея Ч. Бэббиджа состояла в том, что память объёмом 1000 50-разрядных десятичных чисел (по 50 зубчатых колес в каждом регистре) можно использовать для хранения словарей. Ч. Бэббидж привел эту идею в качестве обоснования для запроса у английского правительства средств, необходимых для физического воплощения аналитической машины, которую ему так и не удалось построить [4].

Материальное воплощение идеи машинного перевода появилось лишь в 1947 сразу после появления первых ЭВМ в США. Первая публичная демонстрация машинного перевода в России состоялась в 1954 году. С тех пор началось негласное соревнование между двумя странами в этом направлении.

К середине 1960-х в США для практического использования были предоставлены две системы русско-английского перевода: MARK (в Департаменте иностранной техники ВВС США); GAT (разработка Джорджтаунского университета, использовалась в Национальной лаборатории атомной энергии в Окридже и в центре Евратома в г. Испра, Италия).

Другое направление работ возникло в отделении прикладной математики Математического института АН СССР (ныне ИПМ им. М. В. Келдыша РАН). В этой работе принимали участие А. А. Ляпунов, О. С. Кулагина, Т. Н. Молошная и др., работа велась по машинному переводу математических текстов с французского языка на русский, а также над алгоритмом англо-русского перевода [4].

[1] Погуляев Иван Александрович, студент гр АДб-12-1 Института архитектуры и строительства, e-mail: idipogulyai@mail.ru
Pogulyaev Ivan, a student of group АДб-12-1, The institute of architecture and construction, e-mail: idipogulyai@mail.ru
[2] Игнатьева Елена Павловна, старший преподаватель кафедры иностранных языков для технических специальностей № 2, e-mail: ele20334045@yandex.ru
Ignatyeva Elena, Senior teacher of Foreign Languages for Engineering Specialties № 2 Department of Applied Linguistics Faculty, e-mail: ele20334045@yandex.ru

Первое поколение систем машинного перевода базировалось на алгоритмах последовательного перевода «слово за словом», «фраза за фразой». Возможности таких систем определялись доступными размерами словарей, прямо зависящими от объёма памяти компьютера. Перевод текста осуществлялся отдельными предложениями, смысловые связи между ними никак не учитывались. Такие системы называют системами прямого перевода. На смену им со временем пришли системы последующих поколений, в которых перевод от языка к языку осуществлялся на уровне синтаксических структур. В алгоритмах перевода использовался набор операций, позволяющий путем анализа переводимого предложения построить его синтаксическую структуру по правилам грамматики языка входного предложения (так же, как учат детей языку в средней школе), а затем преобразовать её в синтаксическую структуру выходного предложения и синтезировать выходное предложение, подставляя нужные слова из словаря. Подобные системы называются Т-системами (Т - от английского слова «transfer – преобразование»). Наиболее совершенным считается подход к построению систем машинного перевода на основе получения некоторого, независимого от языков, смыслового представления входного предложения путем его семантического анализа. Затем производится синтез выходного предложения по полученному смысловому представлению. Такие системы называют И-системами (И – от слова «интерлингва»). Считается, что следующие поколения систем машинного перевода будут относиться к классу И-систем.

Уже в 50-е годы XX века А. А. Ляпунов говорил о переводе путем извлечения смысла переводимого текста и его представления на другом языке. Однако такое видение проблемы перевода оказалось в то время слишком смелым.

Даже в современных условиях весьма развитой компьютерной технологии эта задача не решена. Однако некоторые частные результаты, связанные с семантическим анализом текстов, были получены и опубликованы в трудах IFIP (International Federation for Information Processing). В частности был произведён отбор средств, при помощи, которых были построены алгоритмы, эти алгоритмы стали использоваться в дальнейшей работе по переводу. Основная проблема, возникшая при разработке алгоритмов – это подбор словарного состава, который нужно вводить в машину.

Таким образом, потребность в создании теоретических основ машинного перевода привела к формированию нового направления в лингвистике, называемого структурной, прикладной, математической лингвистикой. Формирование этого направления в СССР относится ко второй половине 50-х годов. Ведущую роль в нем сыграли математики А. А. Ляпунов, В. А. Успенский [3].

Нельзя не упомянуть теорию «Текст-Смысл-Текст», созданную И.А. Мильчуком в середине 1960-х годов. По замыслу её создателей, ТСТ является универсальной теорией, то есть может быть применима к любому языку. Теория «Смысл ⇔ Текст» представляет собой описание естественного языка, понимаемого как устройство («система правил»), обеспечивающее человеку переход от смысла к тексту («говорение», или построение текста) и от текста к смыслу («понимание», или интерпретация текста); отсюда символ двунаправленной стрелки в названии теории. При этом приоритет в исследовании языка отдаётся переходу от смысла к тексту: считается, что описание процесса интерпретации текста может быть получено на основе описания процесса построения текста. Теория постулирует многоуровневую модель языка, то есть такую, в которой построение текста на основе заданного смысла происходит не непосредственно, а с помощью серии переходов от одного уровня представления к другому. Помимо двух «крайних» уровней: фонологического (уровня текста) и семантического (уровня смысла), выделяются поверхностно-морфологический, глубинно-морфологический, поверхностно-синтаксический и глубинно-синтаксический уровни. Каждый уровень характеризуется набором собственных единиц и правил представления, а также набором правил перехода от данного уровня представления к соседним. На каждом уровне мы имеем дело, таким образом, с особыми представлениями текста – например, глубинно-морфологическим, поверхностно-синтаксическим и т. п. [2]. Отметим, что именно система «Текст-Смысл-Текст» являлась на тот период более удачной, так как эта система строилась с учётом не только глубинной структуры языка, но и с учётом поверхностной структуры, т. е авторы данной концепции попытались учесть тонкости естественного языка, что является одной из самых сложных задач в создании машинного перевода. Именно эта система впоследствии легла в основу создания системы бейсик, от которой в дальнейшем пошли другие искусственные языки.

Понимание того, что разработка в области машинного перевода – это стратегически важное направление, привело к тому, что в 60-х гг. XX в. началась подготовка кадров в области автоматической переработки текстов на филологическом факультете МГУ, также в Ленинградском и Новосибирском университетах. Под математической лингвистикой понималось изучение языка как абстрактной знаковой системы с целью построения теоретической основы машинного перевода и создания конкретных алгоритмов перевода. В таком понимании математическая лингвистика составляла часть семиотики – общей теории знаковых систем.

Заметим, что в те же годы формальная теория грамматик развивалась в США в трудах Н. Хомского, ставших классическими для области искусственных языков, в частности языков программирования.

Двадцатилетие (1956–1976) один из основателей направления математического анализа В. А. Успенский в своих воспоминаниях назвал «серебряным веком» структурной, прикладной и математической лингвистики в СССР (предположительно, по аналогии с «серебряным веком» русской поэзии) [4].

Первые коммерческие продукты машинного перевода, нашедшие практическое использование, появились в середине 80-х годов. Они были реализованы на персональных компьютерах и являлись системами прямого перевода, возможности которых базировались на огромных (по сравнению с первыми системами) словарях, а не на умении анализировать и синтезировать тексты.

Выделим два принципиально разных подхода к построению алгоритмов машинного перевода: основанный на правилах (rule-based) и статистический, или основанный на статистике (statistical-based). Первый подход является традиционным и используется большинством разработчиков систем машинного перевода (ПРОМТ в России, SYSTRAN во Франции, Linguatec в Германии и др.) Ко второму типу относится популярный сервис Яндекс.Перевод, Переводчик Google, а также новый сервис от ABBYY.

Поясним, что статистический машинный перевод – это разновидность машинного перевода текста, основанная на сравнении больших объёмов языковых пар. Языковые пары, содержащие предложения на одном языке и соответствующие им предложения на втором, могут быть как вариантами написания двух предложений человеком – носителем двух языков, так и набором предложений и их переводов, выполненных человеком. Таким образом, статистический машинный перевод обладает свойством «самообучения». Чем больше в распоряжении имеется языковых пар и чем точнее они соответствуют друг другу, тем лучше результат статистического машинного перевода. Под понятием «статистического машинного перевода» подразумевается общий подход к решению проблемы перевода, который основан на поиске наиболее вероятного перевода предложения с использованием данных, полученных из двуязычной совокупности текстов. В качестве примера двуязычной совокупности текстов можно назвать парламентские отчеты, которые представляют собой протоколы дебатов в парламенте. Двуязычные парламентские отчеты издаются в Канаде, Гонконге и других странах; официальные документы Европейского экономического сообщества издаются на 11 языках; а Организация объединенных наций публикует документы на нескольких языках. Как оказалось, эти материалы представляют собой бесценные ресурсы для статистического машинного перевода [1].

С практической точки зрения, имея в виду качество результирующего текста и его соответствие исходному, программы машинного перевода подразделяют на три категории:

1) полностью автоматический перевод;

2) автоматизированный машинный перевод при участии человека;

3) перевод, осуществляемый человеком с использованием компьютера [5].

Так как на сегодняшний день всё ещё не решены проблемы автоматического понимания, перевода и синтеза текстов, то программы машинного перевода первой из названных категорий являются завтрашним днём.

Программы второй категории разработчики называют МТ-программы (от Machine translation – машинный перевод). Реально автоматизированный (с участием человека) машинный перевод возможен только в условиях искусственно ограниченного, как по словарному запасу, так и по грамматике, языка.

Программы третьей категории разработчики называют ТМ-программы (от translation memory – память перевода). Эту категорию программ применяют профессиональные переводчики, осознавшие преимущества автоматизации их работы с помощью компьютеров. Основу ТМ-программ составляют специализированные словари, соответствующие тематике переводимого текста. При переводе используются конструкции и значения слов и устойчивых словосочетаний, выбранные профессиональным переводчиком, и занесенные в словари системы, а полученный текст подвергается интенсивному редактированию. Словари и уже переведенные фрагменты текстов, запоминаемые в ТМ-системе, могут быть повторно использованы в больших коллективных проектах, ими можно обмениваться. Поэтому ТМ-системы представляют собой важное средство автоматизации труда профессиональных переводчиков.

Итак, мы проанализировали, как идея машинного перевода трансформировалась в конкретный продукт, пока не идеальный, но с конкретными очертаниями. На развитие идеи и получение результатов понадобилось почти 200 лет. Но именно последние 50 лет стали поистине решающими для лингвистической науки в области машинного перевода. За это время, сменилось несколько поколений систем машинного перевода – от первых программ, использовавших ограниченные ресурсы универсальных компьютеров первого поколения – до современных коммерческих продуктов, использующих

мощные ресурсы серверов и персональных компьютеров, включая ПК, в которых можно размещать карманные словари, а также компьютерные сети.

По мере снятия технических ограничений, налагаемых возможностями компьютеров по производительности и памяти, становилось ясно, что проблема перевода текста с одного естественного языка на другой принципиально не сводится только к перекодировке слов. Главная задача – представление контекста, смыслового содержания переводимого текста, знаний о понятиях предметной области, к которой относится переводимый текст. И, безусловно, сложнейшей задачей остаётся создание системы для перекодировки устной речи. Современные достижения в области вычислительной техники, информационных технологий и технологий телекоммуникаций позволят решить эти задачи в обозримом будущем.

Таким образом, технология машинного перевода много лет является важным и перспективным звеном в лингвистике. Системы машинного перевода ещё далеки от совершенства, учёные всего мира работают над улучшением технологий перевода. Их исследования помогают разрушить языковой барьер и улучшают жизнь всех жителей Земли.

**Библиографический список**

1. Машинный перевод: википедия [Электронный ресурс]. – Режим доступа: wikipedia.org (дата обращения: 15.08.15).

2. Круассанов С.П. Теория «Смысл ⇔ Текст» [Электронный ресурс]. – Режим доступа: www.philol.msu.ru/~otipl/new/mscl/2007/example.pdf

3. Рейтблат А.И. Комментарий в эпоху интернета // Новое литературное обозрение. – 2004. – № 66. – С. 82–90.

4. Филинов Е.Н. История машинного перевода [Электронный ресурс]. – Режим доступа: www.computer-museum.ru (дата обращения: 30.04.15).

5. Тараскин А.А. Машинный перевод [Электронный ресурс]. – Режим доступа: www.study – english.info (дата обращения: 30.09.15).